Universidade Estadual de Campinas
Faculdade de Tecnologia

Daniel Bezerra Barros

LOCALIZAÇÃO DE FONTE DE CONTAMINAÇÃO EM CASO DE INTRUSÃO QUÍMICA EM REDES DE DISTRIBUIÇÃO DE ÁGUA

Limeira
2020

Daniel Bezerra Barros

LOCALIZAÇÃO DE FONTE DE CONTAMINAÇÃO EM CASO DE INTRUSÃO QUÍMICA EM REDES DE DISTRIBUIÇÃO DE ÁGUA

Dissertação apresentada à Faculdade de Tecnologia da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Mestre em Tecnologia, na área de Ambiente.

Orientadora: Prof.ª Dr. ªLubienska Cristina Lucas Jaquiê Ribeiro
Coorientador: Prof. Dr. Bruno Melo Brentan

Este exemplar corresponde à versão final da dissertação do aluno Daniel Bezerra Barros, e orientado pela Prof.ª Dr. ªLubienska Cristina Lucas Jaquiê Ribeiro.

Limeira
2020

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca da Faculdade de Tecnologia
Felipe de Souza Bueno - CRB 8/8577

Barros, Daniel Bezerra, 1992-
B278L Localização de fonte de contaminação em caso de intrusão química em redes de distribuição de água / Daniel Bezerra Barros. – Limeira, SP : [s.n.], 2020.

Orientador: Lubienska Cristina Lucas Jaquiê Ribeiro.
Coorientador: Bruno Melo Brentan.
Dissertação (mestrado) – Universidade Estadual de Campinas, Faculdade de Tecnologia.

1. Abastecimento de água. 2. Contaminantes. 3. Aprendizado de máquina. I. Ribeiro, Lubienska Cristina Lucas Jaquiê, 1977-. II. Brentan, Bruno Melo, 1990-. III. Universidade Estadual de Campinas. Faculdade de Tecnologia. IV. Título.

Informações para Biblioteca Digital

Título em outro idioma: Location of contamination source in case chemical intrusion in water distribution network
Palavras-chave em inglês:
Water-supply
Contaminants
Machine learning
Área de concentração: Ambiente
Titulação: Mestre em Tecnologia
Banca examinadora:
Lubienska Cristina Lucas Jaquiê Ribeiro [Orientador]
André Luis Sotero Salustiano Martim
Victor de Barros Deantoni
Data de defesa: 16-01-2020
Programa de Pós-Graduação: Tecnologia

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0000-0002-3664-8701
- Currículo Lattes do autor: http://lattes.cnpq.br/7433847496161630

FOLHA DE APROVAÇÃO

Abaixo se apresentam os membros da comissão julgadora da sessão pública de defesa de dissertação para o Título de Mestre em Tecnologia na área de concentração de Ambiente, a que submeteu o aluno Daniel Bezerra Barros, em 16 de janeiro de 2020 na Faculdade de Tecnologia – FT/UNICAMP, em Limeira/SP.

Profa. Dra. Lubienska Cristina Lucas Jaquiê Ribeiro
Presidente da Comissão Julgadora

Prof. Dr. André Luis Sotero Salustiano Martim
FEC/UNICAMP

Prof. Dr. Victor de Barros Deantoni
PUC/CAMPINAS

Ata da defesa, assinada pelos membros da Comissão Examinadora, consta no SIGA/Sistema de Fluxo de Dissertação/Tese e na Secretaria de Pós Graduação da FT.

,

''Para ser grande, sê inteiro: nada
Teu exagera ou exclui.
Sê todo em cada coisa. Põe quanto és
No mínimo que fazes.
Assim em cada lago a lua toda
Brilha, porque alta vive.''

Fernando Pessoa

Agradecimentos

Um dia eu escolhi fazer mestrado, e como consequência viria a mudança de estado, a vida longe da família e a volta à vida acadêmica. Muitas coisas pesavam contra, mas minha decisão foi abraçada por minha família e segui com a escolha.

Fui apresentado a professora Lubienska, cuja proposta era receber um menino do interior, desconhecido, sem nenhuma experiência e num tema de pesquisa que estava se iniciando. Ela me aceitou e me tratou como filho desde o primeiro dia em que nos conhecemos. Me introduziu nos caminhos da vida acadêmica, me mostrou um mundo até então desconhecido. E se tornou meu maior exemplo de professor.

O grupo de pesquisa já tinha dois membros, Eva e Sandra, as quais se tornaram amigas de lutas, vitórias, derrotas e devaneios, tanto acadêmicos quanto pessoais. Meninas com os corações enormes, acolhedores e dispostos. Mesmo com os prazos sempre curtos e as infindáveis programações e estudos, nós dividimos momentos de risadas e de cafés, de dúvidas e de certezas, de otimismos e pessimismo, de sonho e realidade, de conversa e silêncio, e nunca momentos de guerra.

Tive como coorientador o Bruno, rapaz novo, com um currículo excelente e que teria que parar para me ensinar programação básica. Com ele partilhei batalhas, caminhadas, dúvidas, sonhos e lanches. Ele me ensinou, me corrigiu, me repreendeu e me orientou. Ele se tornou meu amigo e meu maior exemplo de pesquisador.

Partilhei um apartamento e a vida com a Samira, amiga de faculdade e da vida. Com ela conheci a cidade grande e os problemas de uma casa. Com ela compartilhei sonhos, dores, alegrias, vitórias e problemas pessoais. Com ela passei momentos inesquecíveis.

Vivi longe da família, perdi momentos de comunhão, mas sempre senti o amor deles por mim. Mesmo vivendo distante fui amparado por meus irmãos, ajudado pelos meus cunhados, abraçado pelos meus sobrinhos e amado por meus pais. “Afinal, aquilo que amamos sempre será parte de nós”.

Segui rumo ao desconhecido, mas sempre Deus zelou por mim.

Hoje eu sei que fiz uma boa escolha.

Resumo

O sistema de abastecimento de água pode ser visto como uma estrutura física em potencial para ações de ataques terroristas por estar vulnerável a uma gama de ameaças, dentre elas a contaminação química. Investigar as possíveis ameaças destas contaminações no sistema de abastecimento de água, por meio da detecção precoce de eventos e ações pode ajudar as concessionárias a tomarem decisões assertivas em caso de o sistema estar vulnerável a essas ações. Assim, há a necessidade do uso de ferramentas para o auxílio na identificação desses eventos e ações. Esta pesquisa desenvolveu ferramenta computacional capaz de diminuir o empirismo na tomada de decisão e minimizar o número de vítimas em caso de ataques em sistemas de água potável, através de um modelo de localização de intrusão de contaminação por compostos químicos. Em uma primeira fase, com o propósito de localizar anomalias nos dados de monitoramento por sensores, foram utilizadas Redes Neurais, Máquinas de Vetores de Suporte e Random Forest aplicadas na classificação de sinais de concentração de cloro em uma rede com topologia simplificada. Em uma segunda fase, o algoritmo de classificação Random Forest, devido ao seu melhor desempenho, foi aplicado em uma rede de distribuição com maior complexidade e outras abordagens de *output* para modelar a resposta do algoritmo. Com a pesquisa foi possível aumentar a análise do risco adverso ao sistema com tecnologias adequadas e compatíveis para lidar com as ameaças que comprometem a qualidade da água no sistema.

Palavras-chave: Abastecimento de Água; Contaminante; Aprendizado de Máquinas.

Abstract

The water supply system can be seen as a potential physical structure for terrorist attack actions because it is vulnerable to a range of threats, including chemical contamination. Investigating the potential threats of such contamination in the water supply system through early detection of events and actions can help utilities make assertive decisions if the system is vulnerable to such actions. Thus, there is a need to use tools to assist in identifying these events. This research developed a computational tool capable of reducing decision-making empiricism and minimizing the number of casualties in case of attacks on drinking water systems, through a model of intrusion location of chemical contamination. In the first phase, in order to locate anomalies in the sensor monitoring data, we used Neural Networks, Support Vector Machines and the Random Forest algorithm applied to the classification of chlorine concentration signals in a network with simplified topology. In a second phase, the Random Forest classification algorithm, due to its better performance, was applied in a more complex distribution network with new output approaches and obtained satisfactory results. The research made it possible to increase the analysis of adverse system risk with appropriate and compatible technologies to deal with threats that compromise water quality in the system.

Keywords: Water-Supply, Contaminants, Machine Learning.

Lista de Figuras

Lista de Tabelas

Lista de Siglas, abreviaturas e símbolos

ANVISA – Agência Nacional de Vigilância Sanitária

BWSN – The Battle of the Water Sensor Networks

DLL - Dynamic Link Library

JYN - Rede de Distribuição - Jiayou

MLP - Multilayer Perceptron

MSX - Multi-Species Extension

NSGA-II – Algoritmo genético multiobjetivo

PA – Paration

PSO - Particle Swarm Optimization

RDA – Rede de Distribuição de Água

RZM - Rede de Distribuição Al-Zahrani e Moied

RF - Random Forest

RN - Redes Neurais Artificiais

RNR - Redes Neurais Recorrentes

SDA – Sistema de Distribuição de Água

SVM - Support Vector Machine

UFC – Unidade de formação de colônias

uT – Unidade de Turbidez

Sumário

1. INTRODUÇÃO

Apesar da preocupação constante dos pesquisadores para o desenvolvimento de novas tecnologias, métodos, técnicas e/ou ferramentas para o aumento da confiabilidade dos sistemas urbanos, garantir a segurança continua sendo uma tarefa árdua. É grande a quantidade de ameaças que a população está submetida, sendo o terrorismo uma das ações com alto impacto e que recebe muita atenção nas mídias de informação (YAGHLANE e AZAIEZ, 2016). Aly e Green (2010) apresentaram várias definições sobre o terrorismo, dentre elas está a definição considerada pela Organização das Nações Unidas que descreve o terrorismo como "*conjunto de atos criminosos contra um Estado destinado ou calculado a criar uma forma de terror na mente das pessoas, em particular, de grupos ou público em geral*" (GA, 1995). Outra definição considerada é a dos Estados Unidos da América que descreve o terrorismo como "*conjunto de ações que compreende o uso ilegal da força ou da violência contra pessoas ou bens para intimidar ou coagir, visando à consecução de objetivos políticos ou sociais*" (STAFF, 2013).

Consequências diversas podem surgir de ações terroristas, como bombardeios, sequestros, ataques químicos e cibernéticos entre diversos outros métodos utilizados por pessoas e/ou organizações (NEUMANN, 2009). Esses temas são constantemente estudados e debatidos por especialistas, no entanto, um assunto que recebe pouca atenção e pode causar perturbações e mortes é o terrorismo pela água, em que o sistema de abastecimento é ameaçado pela intrusão de contaminação química e biológica, ataques cibernéticos nos controles do sistema e a sabotagem físicas nas instalações (PING, 2010).

As redes de abastecimento de água apresentam diversos pontos vulneráveis, como reservatórios e hidrantes, uma grande abrangência e variados pontos de acesso, tornando a identificação de uma contaminação rede, acidental ou intencional, um processo complexo e tema de pesquisas atuais (LAIRD *et al*., 2005). Pesquisas recentes dissertam sobre a aplicação de sensores que monitoram a qualidade da água em tempo real, comparando a variação da concentração de cloro, turbidez, pH, condutividade ou outros parâmetros que possibilitem o reconhecimento de anomalias na qualidade da água (CLOETE, MALEKIAN E NAIR, 2016; CRISTO e LEOPARDI, 2008).

Utilizando dados oriundos de sensores de qualidade e métodos computacionais, é possível a identificação dos problemas e a caracterização da contaminação, como: local de intrusão, duração da injeção, composição do contaminante; e com isso, medidas de segurança podem ser realizadas imediatamente, tais como: contenção do contaminante, ações policiais e

aviso a população (SALOMONS e OSTFELD, 2010). Com a aplicação eficiente desses métodos de identificação, a vida útil do sistema, as medidas de resposta e o atendimento ao cliente serão melhorados, garantindo a minimização dos custos de manutenção de todo o Sistema de Distribuição de Água (SDA) (RIBEIRO, 2007).

Segundo Bonfim (2016) é necessário que a empresa gestora faça uma análise dos pontos fracos e das possíveis ameaças à integridade dos sistemas de abastecimento de água, de forma a preparar e antecipar os ataques ou contaminações para implementação de ações de prevenção e de resposta a emergências, bem como a criação de metodologia para o restabelecimento do sistema ao estado inicial com máxima eficiência. As ferramentas precisam estar preparadas para minimizar, dissuadir e neutralizar ações que atentem contra a segurança das pessoas, de forma a enfrentar este tema com medidas científicas adequadas.

Visando contribuir com essas questões, diferentes linhas de pesquisas surgiram utilizando métodos como a utilização de matrizes e gráficos, e também métodos computacionais como simulações hidráulicas, otimizações multimodais, aprendizado de máquinas e inúmeros outros (MUKHERJEE *et al*., 2017; PALLETI *et al*., 2016; AGRESTA *et al*., 2015; OHAR e OSTFELD, 2015). Entre os métodos aprendizado de máquinas, em especial as Redes Neurais (RN), vem sendo utilizado com sucesso na previsão de séries em recursos hídricos (MAIER *et al*., 2010) e hidrológicos (DAWSON e WILBY, 2001) e nas relações entre os parâmetros de qualidade da água (MAIER e DANDY, 1996) identificação das anomalias (OLIVEIRA *et al.*, 2018; MAIER e DANDY, 1996).

Porém, as empresas gestoras dos sistemas de abastecimento de água ainda estão aprendendo a lidar com questões em que há comprometimento da qualidade da água, como possíveis contaminações ocasionais ou ataques terroristas. Estudos vêm sendo desenvolvidos para tentar identificar tais intrusões e nortear os operadores na tomada de decisões assertivas para o controle de setores afetados (BARROS *et al.,* 2019; CARDOSO *et al.,* 2018; YAGHLANE e AZAIEZ, 2016; RASEKH *et al.*, 2016; JANKE *et al*., 2006; ALLMAN e CARLSON, 2005).

Considerando o cenário apresentado, é imprescindível o desenvolvimento de alternativas que possam, de forma automática, detectar a presença de contaminantes na rede de distribuição de água, e através desta detecção, determinar a localização da fonte de contaminação em tempo real ou quase real. Neste sentido, o presente trabalho contribui com o desenvolvimento de ferramenta para tomada de decisão através de medidas de segurança e gerenciamento de setores sob intrusão química. Com base matemático-computacional, esse

trabalho faz uso de métodos de aprendizado de máquinas para a localização de fontes de contaminação química em Redes de Distribuição de Água (RDA).

2. OBJETIVOS

2.1. Objetivo Geral

O objetivo deste trabalho é desenvolver um sistema de localização de fontes de contaminante em sistemas de abastecimento de água, utilizando-se de ferramentas de aprendizado de máquinas.

2.2. Objetivos Específicos

- Simular qualidade da água: criar banco de dados contendo o comportamento dos parâmetros de qualidade da água com cenários de contaminação em todos os possíveis pontos de contaminação com diferentes concentrações de contaminante;

- Identificar padrões de dados: aplicar métodos de aprendizado de máquinas para identificar anomalias nos dados e possível associação destas anomalias às fontes de contaminação;

- Estudar o comportamento do contaminante: analisar os impactos e o espalhamento da contaminação até que seja localizada a fonte de contaminação.

3. FUNDAMENTAÇÃO TEÓRICA

A evolução dos humanos das sociedades primitivas para as sociedades agrícolas e sofisticadas baseou-se na disponibilidade de água para cultivo e sobrevivência datando desde 5000 a.C. (ROGERS, 1997). Grandes civilizações iniciaram-se nas margens de rios, como os Egípcios, Hindus e Mesopotâmicos, tornando-se conhecidos como as civilizações hidráulicas, nas quais os corpos hídricos eram utilizados principalmente para consumo, irrigação e transporte (BIEHL, 2003). Para permitir o desenvolvimento das comunidades afastadas de grandes rios, uma vez que os poços e fontes não eram suficientes para atender a demanda da população, fez-se necessário a utilização de sistemas de condução de água, surgindo os primeiros aquedutos Romanos em 312 a.C. (MONTELEONE, 2007).

Contudo, envenenar e poluir os corpos hídricos, indispensáveis para as comunidades, tornou-se tática militar comumente utilizada. Os Assírios, ainda no século VI a.C., lançavam fungos causadores de náuseas, vômitos e morte nas fontes e poços dos inimigos. Cadáveres também eram frequentemente utilizados, com casos relatados desde o século VI a. C. na Mesopotâmia até 1999 em Angola (KREAMER, 2012). Até os aquedutos Romanos, inicialmente construídos no subsolo por segurança, mas tornando-se símbolos de poder imperial e em seguida construídos na superfície, foram alvos de ataques pelos inimigos germânicos, no qual, os cercos mais bem-sucedidos causaram a interrupção do fornecimento de água (ASSANTE, 2009).

Ameaças à interrupção, sabotagem ou envenenamento dos corpos hídricos não estão sujeitas apenas em situações de guerras. O terrorismo definido pelo Departamento de Segurança dos Estados Unidos da América como sendo “o uso de violência ilegal calculada para incutir medo e coagir governos ou sociedades em busca de objetivos políticos, religiosos e ideológicos" (STAFF, 2013) são ameaças constantes em diversos setores a todos os países em que existam conflitos humanos. No entanto, o terrorismo pela água recebe pouca atenção entre os estudiosos e formuladores de políticas, no qual seus efeitos podem provocar grandes perturbações e morte da população, principalmente pela alta vulnerabilidade e a baixa priorização dos governos à segurança dos sistemas de abastecimento de água (PING, 2010).

Outros fatores também podem comprometer a qualidade da água. Baixas pressões nas tubulações podem permitir a entrada de contaminantes por pontos de vazamentos, conexões e durante os reparos (ABDULSHAHEED *et al*., 2017). Características da RDA ainda podem

contribuir com a diminuição da qualidade da água, como a corrosão dos tubos, perdas de desinfetantes, reações indesejáveis, flutuações de turbidez, idade da água na tubulação e práticas inadequadas de construção, manutenção ou reparo (HAJIBANDEH e NAZIF, 2018)

A qualidade da água fornecida aos cidadãos também deve ser um fator preocupante entre os governos, visto que a maioria da população disponibiliza apenas do sistema de distribuição de água para seu consumo. Estima-se que em 2004, 1.6 milhões de mortes e 4% da carga global de doenças, estavam relacionadas ao abastecimento de água e saneamento inseguros, e estes percentuais são distribuídos em maiores escalas entre os países em desenvolvimento (BROWN *et al*., 2013; HUTTON *et al*., 2007; WRIGHT *et al*., 2004). A qualidade é comprometida por agentes químicos e biológicos, com doenças relacionadas a metais pesados, nitratos, orgânicos e outros produtos químicos, as maiores ocorrências estão associadas a infecções biológicas, como cólera, febre tifóide, disenteria amebiana, poliomielite, hepatite A e vários tipos de diarréias, causadas pela falta de higiene pessoal e ingestão ou manuseio de água contaminada por fezes de animais e microrganismos patogênicos (AKO *et al.,* 2009; CLASEN *et al*., 2006).

3.1. Padronização de Potabilidade

Com o intuito de melhorar a qualidade da água, vários métodos de tratamento foram desenvolvidos. A pesquisa de Clasen *et al.* (2006) agrupou as principais categorias de tratamentos, dividindo-os entre a desinfecção por produtos químicos, na qual, comumente, utiliza-se cloro; a remoção física de patógenos por meio de filtração e sedimentação; a desinfecção por calor, com o uso de lâmpadas ultravioletas e do sol; e a combinação entre as categorias. Embora esses métodos sejam efetivos no controle de patógenos microbianos, outro fator tornou-se dilema entre os pesquisadores: a formação de subprodutos de desinfecção, resultantes da reação entre os variados desinfetantes químicos, muitos nos quais são comprovadamente cancerígenos e que são agravados pela alta dosagem de desinfetantes utilizados nos tratamentos (LI, 2008).

Em busca de padronizar a qualidade da água destinada à população, os governos adotaram resoluções determinando parâmetros mínimos e máximos permissíveis na composição da água. No Brasil, a Portaria n° 5 de 2017 do Ministério da Saúde, determina os padrões microbiológicos, organolépticos, químicos e de radioatividade de toda água potável distribuída individual ou coletivamente. A fiscalização deve ser promovida e acompanhada

pelas diferentes escalas de poderes. Nesta portaria, obriga-se a manutenção de determinados parâmetros nos sistemas de distribuição de água, que são:

- Não ultrapassar 500 UFC/mL de bactérias heterotróficas;
- Ausência de coliforme termotolerantes em 100 ml;
- Limite máximo a 5,0 uT para águas subterrâneas com desinfecção, 0,5 uT para filtração rápida e 1,0 uT para filtração lenta;
- Mínimo de 0,2 mg/L para dióxido de cloro e cloro residual livre;
- Máximo de 2,0 mg/L de cloro residual combinado;
- Limite entre 0,6 e 1,7 para fluoretos em função das médias de temperatura;
- Manter pH entre 6 e 9,5.

Esses parâmetros devem ser preservados desde os reservatórios e por toda a extensão da rede até as ligações prediais. A portaria, ainda, apresenta os processos de amostragem para o monitoramento, que variam em função da localização da retirada de amostra, da população abastecida e do tipo de manancial utilizado. A amostragem na saída do tratamento da água limita-se a uma amostra para cada 50 mil habitantes e no sistema de distribuição de água também é variável de acordo com o número de habitantes abastecidos, no qual é retirada uma quantidade de água e analisada em laboratórios. A frequência de monitoramento na saída do tratamento para mananciais superficiais, é de 2 horas para cor, turbidez, cloro, pH e fluoretos, para mananciais subterrâneos, ocorre semanalmente para cor e a cada duas semanas para os demais. No entanto, a frequência de monitoramento no sistema de distribuição de água é mensal para cor, turbidez e cloro, dispensada as análises para pH e fluoretos.

3.2. Métodos de monitoramento

Os resultados do monitoramento dos sistemas de abastecimento de água emitidos pelos laboratórios são demasiadamente lentos, não garantindo a conformidade dos padrões de qualidade exigidos e a segurança dos usuários em tempo real. Este atraso na detecção de contaminação, deliberada ou acidental, leva a consequências graves para saúde pública e nas aplicações de medidas de segurança. Para resolução dessa problemática, nos últimos anos, ocorreram avanços significativos em tecnologias para monitoramento on-line e sistemas de alertas em tempo real ou quase real (ZULKIFLI *et al*., 2017; AISOPOU *et al*., 2012; STOREY *et al.,* 2011).

A implantação de tecnologias e/ou estratégias de monitoramento envolvem uma variedade de sensores biológicos, microfluídicos, químicos e espectroscópicos portáteis, que são conectados a um sistema central de controle, em que esses sensores coletam, integram, analisam e comunicam qualidades incomuns ou incidentes de contaminação. Esses sensores requerem quantidades microscópicas de fluidos (10^{-9}L até 10^{-18} L) e possibilitam a análise no interior da tubulação (ZULKIFLI *et al.,* 2018; ANTONY *et al*., 2014; HALL *et al.*, 2007).

Os biossensores utilizam organismos vivos, como algas e moluscos, para detecção de mudanças súbitas na qualidade da água, porém seu uso na RDA é limitado devido à baixa resistência a desinfetante e produtos químicos. Já os sensores microfluídicos, permitem a detecção de variações nas propriedades físicas e químicas de um fluido, porém são desenvolvidos para detecções específicas, como temperaturas, gases e cloros. Por sua vez, sensores químicos são usados em larga escala, por fazer análises rapidamente, por utilizar instrumentalização online e por possuírem alta sensibilidade a variações da maioria dos parâmetros de qualidade, como cloro livre, turbidez, pH, temperatura, oxigênio dissolvido, cloretos, amônias, nitratos e carbonatos orgânicos (CLOETE, MALEKIAN E NAIR, 2016; ANTONY *et al.,* 2014; HALL *et al.*, 2007).

O monitoramento em todos os pontos da RDA seria dispendioso, devido ao alto custo dos sensores e da mão de obra para instalação e manutenção. Consideram-se, ainda, os riscos de vazamentos acidentais consequentes das tubulações extras necessárias para a alocação dos sensores (CARBONI *et al*., 2016). Portanto, a alocação ideal desses dispositivos tornou-se uma questão crucial, principalmente se considerado as grandes dimensões das redes, com dezena de milhares ou até milhões de nós (KRAUSE *et al*., 2008).

3.3. Alocação de Sensores

As estratégias para o posicionamento de sensores podem ser caracterizadas por abordagens técnicas e abordagens computacionais. A pesquisa de Hart e Murray (2010) estratificou essas propostas em três importantes categorias: opinião de especialistas, métodos de classificação e otimização.

Métodos que consideram a opinião dos especialistas são guiados pelo julgamento humano, como a pesquisa de Dawsey *et al.* (2005), que utiliza uma estrutura que representa as relações casuais entre eventos e observações chamado Rede de Crença Bayesiana. Já os

métodos de classificação utilizam informações adicionais às opiniões dos especialistas, como o uso de ferramentas para análise topográfica do sistema e informações hidráulicas (GHIMIRE E BARKDOLL, 2008; KIRMEYER, 2002).

As técnicas de otimização buscam computacionalmente a alocação dos sensores, e podem considerar, em conjunto, diversos métodos e ferramentas para a tomada de decisão. São capazes de ponderar informações hidráulicas, espaciais, metodologias de análises, formulações matemáticas, atenderem objetivos diversos e, ainda, simular contaminações, incidentes e qualidade (HART e MURRAY, 2010).

3.3.1. Alocação Ótima

Lee e Deininger (1992) foram os primeiros pesquisadores a abordar a alocação de sensores e utilizaram modelos computacionais para análise individual, cujo objetivo era maximizar a cobertura da rede. Nas últimas décadas, houve o crescente desenvolvimento de modelos de otimização e algoritmos não só na maximização da cobertura dos sensores, mas também em métodos que buscam melhorar as informações sobre detecção de eventos, transporte de contaminantes, alarmes de emergência, custos de instalação e manutenção (OSTFELD *et al.*, 2008; PREIS e OSTFELD, 2008).

Em 2006, Ostfeld e Salomons, propuseram uma competição para alocação de sensores, *Battle of Water Sensor Networks* (BWSN), em que foram apresentados os modelos de análises, individuais e múltiplos, utilizados pelos pesquisadores participantes na resolução otimizada de quatro objetivos: o tempo esperado de detecção (Z1), população afetada antes da detecção (Z2), consumo de água contaminada antes da detecção (Z3) e a probabilidade de detecção (Z4).

Em geral, as pesquisas atualmente buscam atender, parcial ou integralmente, esses quatro objetivos, como a pesquisa de Aral *et al*. (2009), que aplica algoritmo genético progressivo na resolução dos quatro objetivos. Já na pesquisa de Cardoso *et al*. (2018), fez-se o uso de algoritmo genético multiobjetivo, cuja finalidade era alocar sensores com o menor tempo de detecção possível e máxima probabilidade de detecção. Ambos utilizados na BWSN, e incluíram um objetivo relacionado a minimização da quantidade de sensores. Mas há, também, pesquisas para a alocação ótima de sensores com funções objetivas próprias, como a pesquisa de Suse *et al*. (2013), que apresentam um modelo matemático-computacional com

conceitos de cobertura de demanda e algoritmos genéticos, cuja a função objetivo é a soma das demandas dos nós cobertas pelas estações de monitoramento.

3.4. Algoritmos bio-inspirados de otimização mono e multiobjetivo

Os algoritmos tornaram-se usuais na resolução de problemas matemáticos com objetivos específicos. Os algoritmos genéticos, utilizados em problemas de otimização, foram apresentados por John Holland na década de 1960, e baseiam-se na evolução e seleção natural das espécies, considerando cruzamentos, mutação, população e estratégias de seleção. Esse método revolucionou as pesquisas de otimização, programação automática, aprendizado de máquinas, economia, sistemas imunológicos, ecologia, genética das populações, evolução e sistemas sociais, e cresce gradativamente com introdução de modificações no algoritmo genético padrão (MITCHELL, 1998).

A natureza também é fonte de inspiração na resolução de problemas complexos, visto que exibe fenômenos extremamente diversificados, dinâmicos, robustos e fascinantes. Os Algoritmos Bio-Inspirados possuem a capacidade de resolver e descrever relações complexas a partir de condições iniciais e regras simples, esses algoritmos tornaram-se ferramentas poderosas para solucionar problemas de otimização, uma vez que os métodos antigos tendem a falhar à medida que há um aumento no tamanho do problema (BINITHA e SATHYA, 2012).

Alguns desses algoritmos são inspirados na vida animal, como o Algoritmo de Enxame de Pássaros (*Bird Swarm Algorithm*) baseado no comportamento das aves (MENG *et al.*, 2016), A Grande Corrida do Salmão (TGRM, do inglês *The Great Salmon Run*) inspirado em peixes (MOZAFFARI e BEHZAPIDOUR, 2012), Colônia de Abelhas (*Bee Colony-Based*) baseados em grupos de insetos (PÉREZ *et al*., 2007) e vários outros. Em geral, possuem tendência em seguir características de métodos de inteligência sociais, como colônias, enxames e cardumes, e são principalmente utilizados em procedimentos de buscas e problemas de seleção de características (DORIGO, 2008). Esses algoritmos destacam-se em pesquisas de otimização aplicadas ao abastecimento de água na alocação de componentes, como sensores (KULKARNI *et al*., 2009) e válvulas (FONTANA *et al*., 2011), aplica-se ainda na operação otimizada de bombas (BRENTAN *et al*., 2018) e válvulas (NICOLINI e ZOVATTO, 2009; CEMBRANO, 2000).

Os algoritmos genéticos multiobjetivo, introduzidos por Fonseca e Fleming em 1993, são uma modificação no nível de seleção dos algoritmos genéticos, e tem como principal atrativo a diversidade de soluções ótimas para a tomada de decisão. Enquanto os algoritmos genéticos com abordagem mono-objetivo analisam os objetivos em uma única vez durante o processo de otimização, os algoritmos multiobjetivos operam com múltiplos objetivos, conflitantes ou não, permitindo envolver minimização e/ou maximização e, ainda, conter restrições nos objetivos e nas variáveis, resultando em um conjunto de soluções que atendem aos diferentes propósitos (DEB, 2014).

O resultado do processo de otimização multi-objetivo é uma fronteira Pareto, que é composta por soluções não-dominadas do problema (REY HORN *et al*., 1993). Para a escolha da solução a ser implementada, avalia-se a importância que cada conjunto de solução representa aos objetivos propostos, sendo, na maioria dos casos, realizada por especialistas. Porém, técnicas automáticas de ajuda ao tomador de decisão estão sendo aplicadas e desenvolvidas nesses processos (CARPITELLA *et al*., 2019; CARDOSO *et al*., 2018a).

Outro método aplicado com sucesso em problemas de processamento de sinais são as Redes Neurais (RN), que possuem a capacidade de se adaptar continuamente a novos dados e rastrear mudanças em um sinal ao longo do tempo (ABRAHAM, 2005). Essas redes são modelos computacionais inspirados nos neurônios naturais e consistem em entradas (sinapses) que são multiplicadas por pesos e calculadas por uma função matemática que ativa os neurônios, ocorre, ainda, a combinação dos dados entre os neurônios para processamento de informações. Esse procedimento oferece melhor desempenho em detecção de padrões robustos, filtragem de sinais, controle adaptativo, análises textuais, otimização, programação e segmentação, compreensão e mineração de dados (BAJPAI *et al*., 2011; DREW e MONSON, 2000).

3.5. Algoritmos de aprendizado de máquinas

Dentre as diversas aplicações de técnicas de aprendizado de máquinas, a modelação de um sistema complexo está entre as mais populares (WU e LIU, 2017; IZQUIERDO *et al*., 2007). Esses métodos de aprendizado de máquinas são aplicados à mineração de dados e possui capacidade de prever novas situações através de informações que descrevem acontecimentos passados. Algumas das principais técnicas de aprendizado de máquinas são os Redes Neurais,

(RN) as Máquinas de Vetor Suporte (SVM, do inglês *Support Vector Machine*) e as Florestas Aleatórias (RF do inglês *Random Forest*) (PAES *et al*., 2018; WITTEN *et al.,* 2016).

As RNs ganharam notoriedade nas pesquisas associadas aos sistemas de distribuição de água, aplicadas no desempenho da estação de tratamento (HAMED *et al*., 2004), na otimização de operação (SALOMONS e*t al.,* 2007), na previsão de demanda (BRENTAN *et al*., 2017; Herrera *et al.*, 2010; JAIN *et al*., 2001), na gestão de pressões (NAZIF *et al*., 2010), na previsão dos parâmetros de qualidade (MAIER e DANDY, 1996), no controle do gasto de energia (AL-ALAWI *et al*., 2007), em simulações hidráulicas (MEIRELLES *et al.*, 2018; RAO e BARMEJO, 2007) e nas áreas que necessitam de previsões, otimizações e análises de dados.

Entre os diversos algoritmos de aprendizado de máquinas, alguns tendem a privilegiar a capacidade de generalização e o mapeamento não linear, como as RN do tipo Perceptron de múltiplas camadas (MLP, do inglês *Multilayer Perceptron*), que possui a capacidade de mapeamento não-linear de dados, as Redes Neurais Recorrentes (RNR), que possibilitam a modelagem de séries temporais, algoritmos que dispõe da junção de técnicas, como as Redes Neurais Auto-regressivas com Entradas Exógenas, (NARX, do Inglês *Non linear Autoregressive with Exogenous Inputs)* que aplica o método MLP e a classe de métodos RNR (ZHE *et al*., 2015) ou ainda as máquinas de vetor suporte que mapeiam espaços característicos em espaços altamente dimensionais a partir de transformações não-lineares.

3.5.1. Rede Perceptron de múltiplas camadas - MLP

As Redes Neurais do tipo MLP, também chamadas de *Feedforward* multicamadas, são modelos de Redes Neurais consagradas na literatura. Esses modelos são ideais para mapeamento e mineração de dados, apropriados para problemas de mapeamento funcional, em que o número de variáveis de entrada afeta as variáveis de saída (ZHANG, 2009).

A arquitetura de uma MLP é variada, mas consiste basicamente em várias camadas de neurônios, como mostra a Figura 1. Essa arquitetura possui uma camada de entrada (*input layer*) que transmite os vetores de entrada (X_i) para a rede, uma ou várias camadas ocultas (*hidden layers*) que processam não-linearmente esses dados de entrada, e finalmente uma camada de saída (*output layer*) que gera vetores de saída ($\hat{y}_t$) (GARDNER e DORLING, 1998).

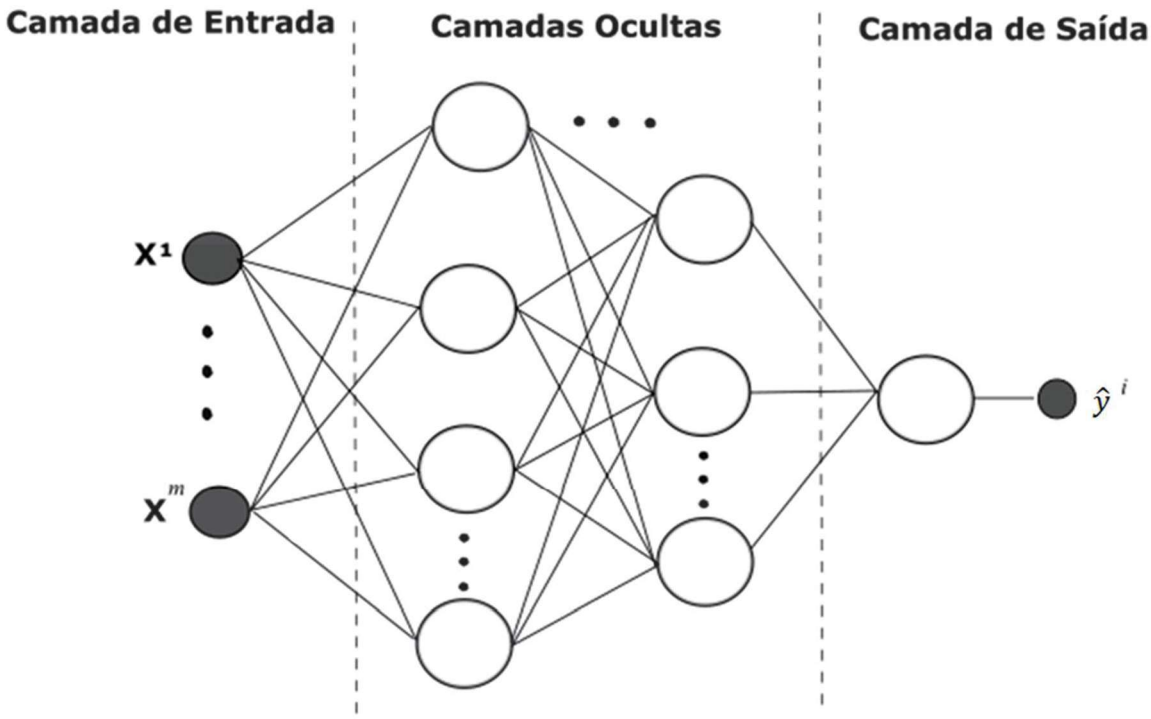

Figura 1 - Arquitetura MLP.

Para uma rede qualquer com apenas uma camada oculta, a saída w^i, será correspondente a um vetor de entrada X^i, dada pela equação 1:

$$f(X_i) = f(\Sigma^{Q}_{q=1} W_q . f(\Sigma^{P}_{p=1} W_p . X^i)) = \hat{y}^i \qquad (1)$$

em que W_q é o vetor de pesos referentes aos neurônios das camadas ocultas q, W_p é o vetor de peso referente ao neurônio da camada de entrada e, X^i é o vetor de dados de entrada i. A ativação neural é dada pela função $f(.)$, ou seja, o quanto um neurônio é afetado por um sinal de entrada. Pode-se observar, ainda, a partir da Figura 1 e da Equação 1 a não linearidade do processamento dos dados, pois os neurônios apresentam ligações com todos os outros das camadas seguintes (GARDNER e DORLING, 1998; SCHUSTER e PALIWAL, 1997).

As MLPs, em geral, são programadas para processar informações e aprender através destas informações, porém é necessário um treinamento desta rede. Esse treinamento envolve o ajuste dos pesos sinápticos da rede para a produção de uma saída específica para o conjunto de entrada com o menor erro possível, mas com a maior generalização possível (SHEPHERD, 2012). O desempenho do treinamento varia entre um determinado problema e as configurações das redes. Para isso, diversos algoritmos de treinamento são disponibilizados, a fim de

solucionar esses problemas, como a Retropropagação (*Backpropagation)* (RUMELHART *et al.*, 1985), Gradiente Conjugado Escalonado (*Scaled Conjugate Gradient)* (MØLLER, 1993), Algoritmo Levemberg-*Marquardt* (MARQUARDT, 1963), e vários outros, foram introduzidos e modificados para resoluções de problemas em diferentes tipos de aplicação (ILONEN *et al.*, 2003).

Esses algoritmos, em geral, são aplicados a problemas com uma quantidade moderada de variáveis. Em problemas com grandes quantidades de dados, o número de algoritmos de treinamento diminui drasticamente, na qual há a necessidade de repartição dos dados para processamento individual ou junções de métodos de treinamento (ILONEN *et al.*, 2003). Uma alternativa que se tornou usual e viável para problemas com muitas variáveis, foi a junção destes métodos de treinamento, como o *Quasi-Newton* (DENNIS e MORÉ, 1977). O mais comumente utilizado, *Levenberg-Marquardt* (MORÉ, 1978), além de se aplicar em redes MLP com muitas camadas ocultas.

O método *Levenberg-Marquardt* apoia-se na aproximação linear de uma função f ao redor de um ponto $\boldsymbol{x}$ de modo que (Equação 2):

$$f(x + \eth x) = f(x) + \nabla f(x) . \eth x \qquad (2)$$

no qual, toma-se como erro de aproximação s entre $f(x^{x})$e $f(x_0)$, de maneira que minimizar a diferença $c - \nabla f(x) . \eth x$ é o mesmo que estimar o ponto ótimo, $x^{\times}$, a partir de x_0, assim na condição de optimalidade, têm-se a Equação 3:

$$\nabla f(x)^{T}. \eth x \nabla f(x) = \nabla f(x)^{T}. c \qquad (3)$$

A fim de obter o valor ótimo, o valor de $\eth x$ pode ser melhorado de modo iterativo, em que dirige para um ponto comum a partir de uma estimativa x_0. O valor de $\eth x$ pode ser estimado a cada iteração, em que é dado pela Equação 4:

$$\eth x = [\nabla f(x)^{T}. \nabla f(x)]^{-1}. f(x) \nabla f(x)^{T}. c \qquad (4)$$

A equação 4 inclui-se no método de Gauss-Newton e possui problemas de convergência, uma vez que a inversão da matriz hessiana, estimada por $\nabla f(x)^{T}. \nabla$, não possui garantia para todos os casos, criando a necessidade da positivação de tal matriz. No intuito de acelerar o

processo de convergência, Levenberg (1944) propôs adicionar uma parcela à estimativa da Hessiana, em que se garanta a positividade da matriz, sendo proposto na Equação 5:

$$ðx = [❑f(x)^T. ❑ f(x) + μI]^{-1}. f(x)❑f(x)^T. c \quad (5)$$

A principal finalidade de μI, é tornar positiva a matriz $❑f(x)^T. ❑$, no qual varia de acordo com as iterações, mas altamente dependente do valor de μ, tornando-se instável em algumas situações. Para reduzir essa instabilidade, Marquardt (1963), apresentou uma formulação alterada, em que há a ponderação de cada componente do gradiente pela sua curvatura, em que há, assim, a Equação 6:

$$ðx = [❑f(x)^T. ❑ f(x) + diag(❑ f(x))μ]^{-1}.f(x)❑f(x)^T. c \quad (6)$$

A grande vantagem desse método está na aproximação da matriz Hessiana como produto do vetor Jacobiano e a positivação da matriz, que se dá pela soma de um valor μ aos autovalores da Hessiana estimada (BRENTAN, 2017).

3.5.2. Máquinas de Vetor Suporte– SVM

As SVM são métodos relacionados a aprendizagem supervisionada, aplicáveis a problemas de classificação e regressão. Nesse método, a solução para o problema depende de um subconjunto de dados de treinamento que são referidos como vetores de suporte (SHMILOVICI, 2009; CAO e TAY, 2003).

Essa classificação utiliza algumas restrições para separar os dados em grupos que possuem características semelhes. Para isso, o método objetiva a produção de um classificador que separe as classes a partir dos dados disponíveis. Pode-se observar na Figura 2, a existência de classificadores lineares para possíveis separações dos dados, mas há apenas um (em verde) que maximiza a distância entre o classificador e o ponto de dados mais próximo de cada classe, esse classificador ideal é denominado de classificador de separação ideal ou hiperplano de separação ideal, quando adiciona-se mais um dimensões os dados (GUNN *et al.*, 1998).

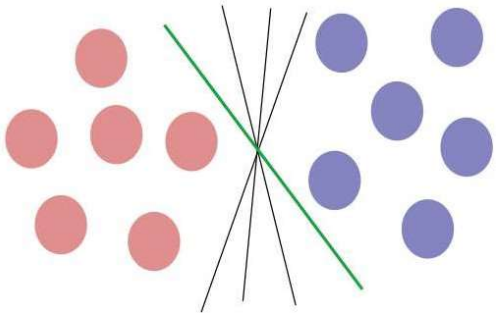

Figura 2 - Separação de dados.

O método SVM classifica a partir de vetores que restringem uma região no espaço, formulados por fronteiras em torno de uma curva de separação, com tamanho s. Em alguns casos, há a necessidade de uma transformação nos dados, afim de encontrar uma dimensão em que posso ocorrer a separação destes dados. A determinação desses classificadores toma como um conjunto de dados rotulados (entrada e saída conhecidas)$\{(x_1, y_1) \ldots (x_n, y_n)\} < X \times R$, Como exemplificado, o objetivo é encontrar o hiperplano de separação ótimo, escrito como:

$$\hat{f}(x) = (w, \$(x)) + b \qquad (7)$$

em que, $\$(x)$ corresponde à transformação dos dados de entrada para o espaço característico, sendo $\$$ a função de mapeamento implícito e x os dados de entrada. Desvia-se no máximo s de um dado observado y_i, que se minimiza simultaneamente a complexidade do modelo, assim os parâmetros w e b são restringidos em conjunto com o problema, definido pela equação 8 e 9:

$$\frac{1}{2}||w||^2 \qquad (8)$$

$$s.\, a|y_i - (w, \$(x)) - b| \leq s \qquad (9)$$

Assume-se que para todo $\hat{f}(x_i)$, existe um valor y_i, que desvia no máximo s. Mesmo assim, fixando as restrições o problema pode não haver solução, sendo necessária a definição de uma fronteira de relaxamento de restrição, em que permite uma maior precisão caso existam valores discrepantes. Essa inclusão das variáveis de relaxamento gera penalidades a desvios maiores que s, permitindo a exploração pelo modelo a todos os pontos existentes na definição dos melhores vetores de suporte. Toma-se $£^+$ e $£^-$ como variáveis de relaxamento, também denominadas de margens brandas, descritos pelas equações 10 e 11.

$$£_i^+ = \hat{f}(x) - y(x_i) \Sigma s \quad (10)$$

$$£_i^- = y(x_i) - \hat{f}(x) \Sigma s \quad (11)$$

por fim, os problemas dos vetores de suporte podem ser ajustados pelas equações 12, 13, 14 e 15:

$$\frac{1}{2}||w||^2 + C\frac{1}{n}\Sigma_{i=1}^{n} (£_i^+ + £_i^-) \quad (12)$$

$$s.a\ y_i - (w, \$(x)) - b \leq s + £_i^+ \quad (13)$$

$$(w, \$(x)) + b - y_i \leq s + £_i^- \quad (14)$$

$$£_i^+, £_i^- \leq 0\ i = 1, \ldots, n \quad (15)$$

no qual *n* é tamanho do lote de treinamento, *C* é o parâmetro relacionado a complexidade do modelo e o erro de treinamento. Esse método é conhecido como sensibilidade s, em que apresenta a tolerância do erro, também conhecido com Hiperparâmetros (BRENTAN, 2017; SHMILOVICI, 2009; CAO e TAY, 2003; CORTES e VAPNIK, 1995).

3.5.3. Random Forest

A utilização do classificador *Random Forest* (RF) ganhou forças nas últimas décadas devido a seu desempenho e sua velocidade de processamento, utilizando predições provenientes de um conjunto de árvores de decisões (BREIMAN, 2001). Este classificador pode ser usado na seleção e classificação de variáveis, discriminando as classes, com poucos ajustes para o alcance de bons resultados. Estas são características importantes, visto que quando se utilizam dados com alta dimensionalidade a seleção das variáveis se torna uma tarefa difícil, propensa a erros e subjetiva (FISCHER e KRAUSS, 2018).

As RF realizam a classificação através de um conjunto de árvores de classificação e regressão (Figura 3). Essas árvores são criadas desenhando um subconjunto de treinamento mediante substituição, ou seja, algumas amostras (X_j) podem ser selecionadas várias vezes, no entanto outras podem não ser selecionadas, há, ainda, divisões internas ($X_j \leq t_k$) que determina a quantidade de ramos que a árvore terá, resultando no indicativo médio das observações (R_i), sendo j a quantidade de variáveis, k a quantidade de ramos e i a quantidade de observações. O

treinamento dessas árvores considera dois terços das amostras e o terço restante é utilizado em um cruzamento interno, em que se estima o desempenho do modelo RF resultante (BELGIU e DRAGUT, 2016).

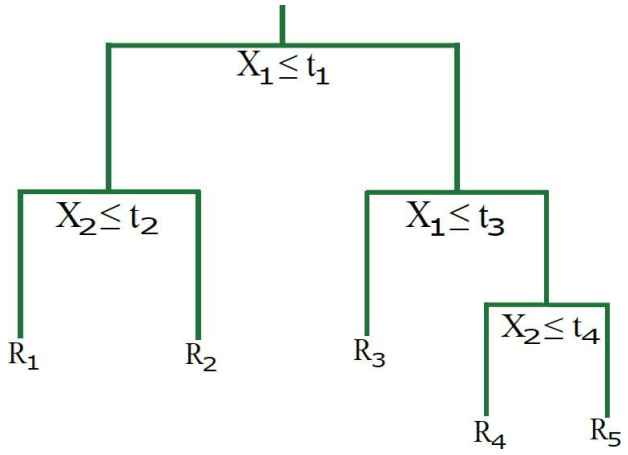

Figura 3 - Exemplo de *Random Forest* (Adaptado, TYRALIS *et al*., 2019).

A aplicação é bastante simples, em que para cada comitê de árvores é extraído uma amostra dos dados originais que podem conter um, mais que um ou nenhum dado dos que são utilizados no treinamento. Em cada comitê, há somente um subconjunto dos recursos disponíveis como possível critério de divisão, parando o crescimento dessa árvore quando uma profundidade máxima é atingida. No fim do processo, o resultado é um comitê de árvores cuja classificação foi determinada por voto majoritário (FISCHER e KRAUSS, 2018; BELGIU e DRAGUT, 2016).

3.6. Simulações hidráulicas e de qualidade

Rossman (1994) desenvolveu o software EPANET, que permitiu a implementação de técnicas computacionais em conjunto com modelos de RDA em ambiente de programação, capaz de analisar o comportamento hidráulico e de qualidade da água em tubulações pressurizadas. O EPANET possui a capacidade de calcular todas as vazões, pressões e níveis de água e qualidade, além da concentração de substâncias, idade da água e o rastreamento da fonte. Este software permite a formação de uma RDA com tubulações, nós, válvulas, bombas, reservatórios de capacidades ilimitadas (RNF) e reservatórios de alturas variadas (RNV) (ROSSMAN, 1999).

Rossman (2000) disponibilizou uma nova versão, EPANET 2.0, que possui interface gráfica e permite a construção de modelos de redes e a possibilidade de importação de

programas de desenhos auxiliados por computador. Os dados do arquivo podem ser editados e visualizados externamente, através de editores de textos, e possui uma extensão definida (.inp). O software, ainda, possui interface integrada com edição dos dados, visualização dos resultados em vários formatos e simulações hidráulicas e de qualidade.

Os conjuntos de ferramentas de cálculos, presentes no software, colaboram na elaboração de pesquisas em todos os componentes da rede, como em melhorias na eficiência dos componentes (BRENTAN *et al*., 2018), controle de pressões (ARAUJO *et al*., 2006), operações de bombas (GIACOMELLO *et al*., 2006), análises de vazões nos tubos (DAS *et al*., 2008), estudos de qualidade (CASTRO e NEVES, 2003), entre outros.

No entanto, esse software realiza a modelação de qualidade considerando apenas uma espécie química. Com o propósito de sanar esta limitação, foi desenvolvida a ferramenta EPANET *Multi-Species Extension* (MSX), que permite a interação de múltiplas espécies químicas, tanto na massa fluida quanto na superfície da tubulação. Para a simulação das interações químicas, o EPANET-MSX, necessita de dois arquivos de entrada (SHANG *et al*., 2008). O primeiro é o arquivo padrão EPANET (.inp), que contém as informações hidráulicas da rede em análise. O segundo arquivo possui as interações dinâmicas dos componentes químicos a serem simulados, como reações e modelos químicos, contendo as informações e as espécies que o usuário deseja obter e possui a extensão (.msx) (SUSE *et al*., 2014). Seus resultados são disponibilizados em arquivo de texto, em formato (.rpt).

O arquivo (.msx) é dividido por sessões sendo a sessão [TITLE] para o título do arquivo; [OPTIONS] para as opções dos valores computacionais, [SPECIES] para nomeação das espécies químicas; [TERMS] os termos utilizados nas expressões químicas. Em pesquisas de contaminação de SDA, algumas sessões possuem maior aplicabilidade, como a sessão denominada [SOURCES], onde se identificam as fontes para a entrada de compostos químicos; a sessão [QUALITY] que fornece as condições iniciais de qualidade; e a sessão [PATTERNS] que define os padrões de tempo para aplicação dos compostos químicos (SHANG *et al*., 2008).

Esse conjunto de ferramentas de software pode ser integrado a programas autônomos ou a bibliotecas de funções, o qual possibilita a personalização de aplicativos e softwares (SHANG *et al*., 2008).

3.7. Incorporação ao ambiente de programação

A pesquisa de Eliades *et al*. (2016) introduziu um software de código aberto EPANET-MATLAB Toolkit, sob a Licença Pública da União Europeia (EUPL), uma biblioteca *wrapper* que possibilita a chamada e execução das funções do EPANET em linguagem de programação C++ em interativo ao MATLAB®. Nesse ambiente de programação, o usuário dispõe de todas as informações da RDA elaborada no EPANET, no qual possibilita modificar e criar redes, executar simulações entre múltiplas espécies químicas e a visualização, em si, da rede, proporcionando maior liberdade no processamento e modificação dos dados (ELIADES *et al*., 2016; SANZ e PÉREZ, 2013).

Essa liberdade de processamento permitiu a incorporação de técnicas computacionais nas análises, predições, melhorias e otimizações das RDA. Como Cardoso *et al*. (2018) que fazem uso da ferramenta em conjunto com algoritmo multiobjetivo na alocação de sensores de monitoramento. E, por fim, Oliveira *et al*. (2018), que empregam um processo estatístico utilizando redes neurais artificiais, para gerar sinais de alerta em possíveis mudanças na qualidade da água, a partir da comparação entre as séries temporais estimadas e as medidas.

A pesquisa de Cristo e Leopardi (2010) utiliza o software EPANET para simular uma contaminação acidental na RDA e, através dos dados gerados na simulação, aplica-se a abordagem de cobertura de demanda para selecionar possíveis nós fontes e uma otimização discreta para inversão do caminho de qualidade da água.

Já a pesquisa de Liu *et al*. (2011) localiza a fonte de contaminação através das observações de um conjunto de sensores de qualidade alocados na RDA. Os autores utilizam uma técnica de otimização dinâmica adaptativa para avaliar e identificar soluções com os erros de predição minimizados e um método de busca não linear baseado em algoritmos evolutivos na determinação da fonte.

Laird *et al*. (2005) apresentam um algoritmo de rastreamento de origem de contaminação para identificar a hora e a fonte de contaminação. Esse método é eficiente para grandes redes, pois reformula as expressões dos tubos com um conjunto de restrições de atraso de tempo, retirando a necessidade de discretização ao longo do comprimento do tubo, em que são requeridas as concentrações apenas nos limites dos tubos ou nos nós da RDA.

4. APRENDIZADO DE MÁQUINAS PARA LOCALIZAÇÃO DE FONTES DE CONTAMINAÇÃO

Para a aplicação da metodologia proposta é necessário um banco de dados que contenha as interações dos componentes químicos no sistema de distribuição e que haja cenários de contaminação em todos os possíveis pontos de intrusão. Esse banco de dados será gerado através de simulações hidráulicas e de qualidade. Após a conclusão do banco de dados, aplicam-se os métodos de aprendizado de máquinas para identificar anomalias nos dados e associá-las às fontes de contaminação.

4.1. Simulação da qualidade da água

Para a criação do banco de dados, utilizou-se o EPANET-MSX (SHANG *et al*., 2008) e o software EPANET 2.0 (ROSSMAN, 2000) integrado ao ambiente de programação MATLAB®, que é possível através do EPANET-MATLAB Toolkit (ELIADES *et al*., 2016). Primeiramente é necessário a escolha da rede de abastecimento a ser estudada e a construção do arquivo de entrada para a simulação hidráulica e posteriormente, a construção do arquivo (.msx) para a rede em estudo, que contém as informações químicas a serem analisadas. Para isso, foram utilizadas as pesquisas de Cardoso *et al.* (2018), Ohar *et al*. (2015) e Oliveira *et al*. (2018), nos quais utilizam o pesticida Paration em métodos de alocação de sensores e para detecção de contaminação química na rede, respectivamente.

A pesquisa de Cardoso *et al*. (2018) apresenta uma metodologia para alocação de sensores de qualidade no SDA, seguindo a aplicação do pesticida Paration em interação com o cloro, no qual se aplica o Algoritmo Multiobjetivo NSGA-II. A aplicação desta técnica faz-se necessária, visto que as leituras dos sensores de qualidade, alocados de maneira ótima, permitem a análise do comportamento químico no interior da tubulação. Cardoso *et al*. (2018) utiliza três funções objetivos para a alocação dos sensores: minimizar o tempo de detecção, maximizar a probabilidade de detecção e minimizar a quantidade de sensores.

No presente trabalho consideraram-se dois destes objetivos: minimizar o tempo limite de detecção e maximizar a probabilidade de detecção, uma vez que as quantidades de sensores são indicadas em pesquisas ou determinadas pelos órgãos gestores da RDA. Para minimizar o tempo de detecção (t_d) (Eq. 16), é considerado o tempo decorrido entre o início da contaminação até a detecção do contaminante por um sensor, quando se obtém a primeira

detecção (t_j) refere-se a localização do sensor j, no qual o t_d é o mínimo entre a rede de sensores (n_c).

$$t_d = \min t_j \ , j = 1, \ldots\ldots, n_c \qquad (16)$$

A função objetivo de minimizar o tempo limite de detecção (Fo_1) considera os eventos que foram que foram detectados, considerando os menores valores de t_d, estimado na Eq. 17:

$$Fo_1 = E(t_d) \qquad (17)$$

A função para maximizar a probabilidade de detecção (Fo_2) considera os maiores valores de cobertura que o conjunto de sensores testado obtém e pode ser expressa pela Eq. 18:

$$Fo_2 = \frac{1}{c} \sum_{r=1}^{c} d_r \qquad (18)$$

em que $d_r = 1$ se o evento de contaminação r for detectado por pelo menos um sensor de monitoramento; S representa a quantidade total de cenários de contaminação (CARDOSO *et al.,* 2018).

A pesquisa de Ohar *et al*. (2015) apresenta detalhadamente as reações dos organofosforados com o cloro livre na água, reações estas que geram um decaimento abrupto na concentração do cloro (Figura 4) quando em contato com compostos organofosforados. Já a pesquisa de Oliveira *et al*. (2018) apresenta as equações estequiométricas e de degradação de um organofosforado, pesticida Paration, em interação com o cloro presente na água, considerando as características brasileiras de potabilidade. Baseou-se nestas metodologias para a construção do arquivo de entrada MSX, por considerar que ocorre uma variação na concentração de cloro quando aplicado diferentes concentrações do pesticida.

As três pesquisas permitem a construção do arquivo de entrada (.msx) para a rede a ser analisada, em conjunto com a localização dos sensores de qualidade. Porém, alterações pontuais, relativas à composição química, são necessárias, tanto no arquivo (.msx), quanto no código de simulação. Como as sessões [SOURCES] e [PATTERNS] do arquivo (.msx), passam a ser gerados pelo EPANET-MATLAB Toolkit, há a necessidade de tornar essas sessões vazias no arquivo de entrada.

Dada a particularidade de cada RDA, os pontos de monitoramento devem ser introduzidos no código de simulação. Já na geração dos cenários de contaminação a serem analisados, a concentração, o tempo de início e duração da intrusão do contaminante são escolhidos de tal forma a criar um banco de dados abrangente e com tamanho suficiente para a aplicação de métodos de aprendizado de máquinas. Para isso, consideram-se todos os nós da

rede como fonte de contaminação, aplicados e analisados individualmente, incluídos os nós monitorados. Este banco de dados apresenta a interação de todos os compostos químicos analisados, determinados no arquivo (.msx), bem como alcalinidade e pH, com passo de tempo entre as análises de 600 segundos.

Em cada nó há cenários de contaminação variando de acordo a concentração de contaminante, em que o valor mínimo é 0,1 mg/L e o máximo de 19 mg/L. Este intervalo foi escolhido por cobrir a dose letal (13mg/kg) e concentração de solubilidade (12,4 mg/L). O tempo de início e duração da contaminação é padronizado para todos os cenários, nos quais iniciam-se na metade da simulação com duração de 12 horas de intrusão. A quantidade de cenários de contaminação, considerando os limites de concentrações estipulados, varia de acordo a RDA em que se realiza a simulação, uma vez que em redes grandes o processo de simulação é extenso.

4.2. Aplicação de Aprendizado de Máquinas

O processo de localização em tempo real da fonte de contaminação conta com a aplicação do banco de dados em Redes Neurais, no método SVM e no algoritmo RF. As informações usadas nos treinamentos e validações consideram o efeito da dinâmico do comportamento do cloro para cada passo de tempo e a cada iteração dos métodos. Tendo o banco de dados eventos anômalos, a concentração do contaminante e a localização da fonte conhecidos, é possível a aplicação de ferramentas de aprendizado de máquinas para a classificação dos sinais de concentração em função da fonte do contaminante.

Todos os métodos de aprendizado de máquinas serão treinados para cada cenário de contaminação, a partir dos dados da distância entre um reservatório da rede, concentração do contaminante utilizado e a variação da concentração de cloro. Essa variação ocorre devido a reação entre cloro e pesticida Paration, em que acontece uma redução na concentração do cloro, podendo ser observada na Figura 4, no qual expõe o comportamento padrão de concentração (Figura 4a) e o comportamento após a intrusão do pesticida (Figura 4b).

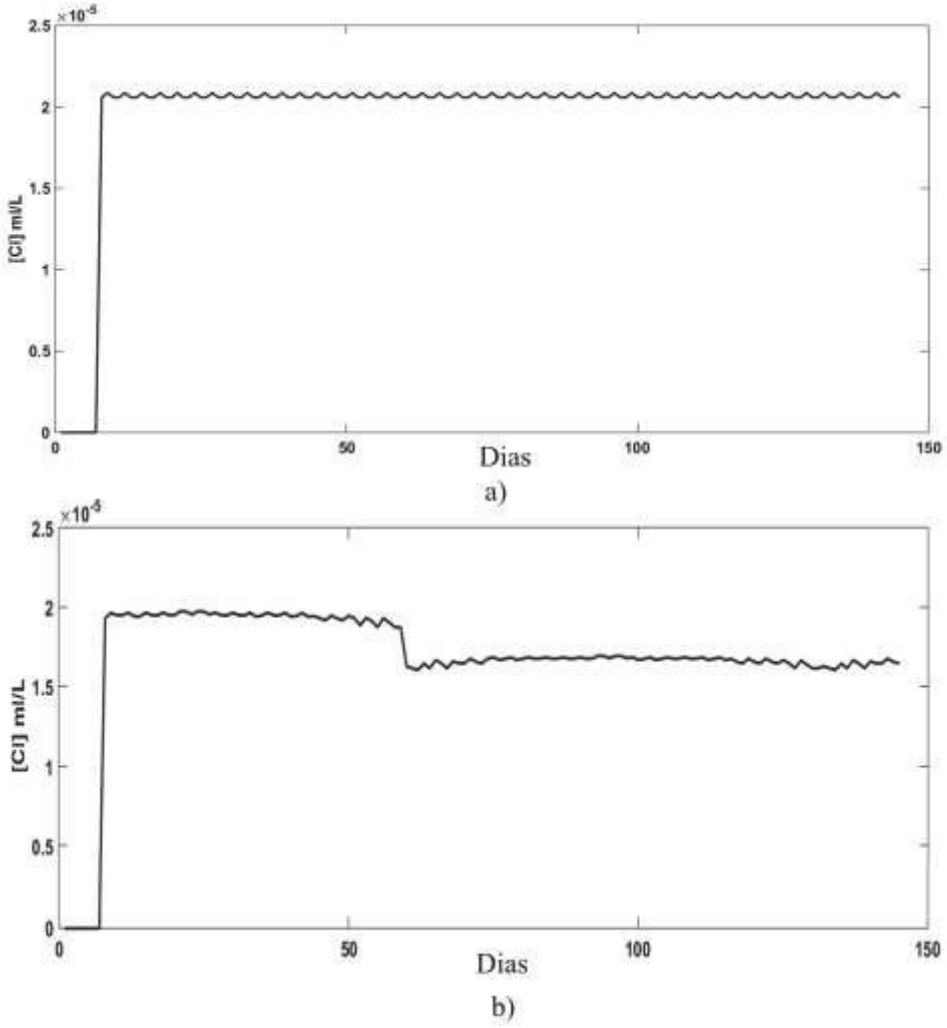

Figura 4 - (a) Comportamento padrão do Cloro. (b) Comportamento do Cloro após a intrusão do pesticida.

Os dados utilizados na camada de entrada (*input*) possuem suas dimensões determinadas pela quantidade de sensores de qualidade e pela quantidade de nós presentes na rede. Para camada de saída (*output*), utilizam-se as informações das distâncias lineares entre um ponto e os possíveis locais de intrusão de contaminante.

4.2.1. Input e Output

Os dados de input utilizados neste trabalho são as concentrações observadas por cada sensor de monitoramento. Conhecidas as concentrações do cloro em cada nó monitorado, no determinado passo de tempo e em qual nó ocorreu a contaminação, poderá ser associado a intrusão de contaminante, uma vez que houver alterações nas concentrações observadas. Há,

também, situações em que a intrusão não afete a concentração do cloro, isto pode ocorrer quando os nós contaminados não são cobertos por sensores.

Esta ligação entre concentração de cloro, passo de tempo e nó contaminado é o tripé fundamental para a localização da fonte em tempo real. Para que a localização em tempo real ocorra, o algoritmo RF será treinado e validado de acordo o passo de tempo em análise, ou seja, o quantidade de treinamentos e validações para cada método será a mesma quantidade de passos de tempo de contaminação simulados. Cada treinamento e validação terão como Input os dados referentes ao passo de tempo em estudo (Figura 5), ocorrendo à inclusão do passo de tempo subsequente, ou seja, o primeiro treinamento e validação contará com os dados relativos ao primeiro passo de tempo, já o segundo treinamento e validação irá dispor dos dados relativos ao primeiro e segundo passos de tempo, assim sucessivamente, sendo o último processo com todos os passos de tempo concatenados.

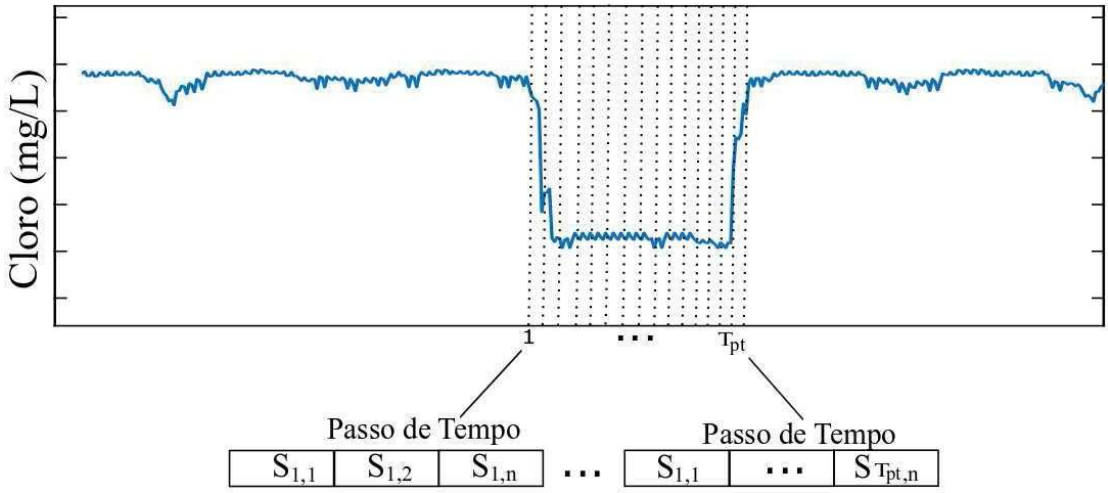

Figura 5 - Inputs Concatenados.

em que n é a quantidade de sensores de monitoramento, S a concentração de cloro observada no passo de tempo e T_{pt} é o total de passos de tempo. A Figura 5 expõe a concentração de cloro, observada por um sensor de qualidade, quando ocorre uma contaminação por Paration, as linhas pontilhadas representam um vetor de entrada para cada passo de tempo que serão concatenados no processo de aprendizado de máquinas.

A abordagem de output conta com as distâncias lineares entre um ponto específico do mapa e os nós utilizados como fontes de contaminação, seguindo a metodologia proposta por Barros *et al*. (2019), no qual utilizam as coordenadas dos nós para calcular as distâncias entre um ponto específico no mapa. Um reservatório será utilizado como ponto base para o cálculo

das distâncias, resultando em um vetor de Output em que há a indicação da distância do reservatório e cada nó contaminado.

4.2.2. Métodos de AM propostos

Foram utilizadas duas redes neurais do tipo MLP, ambas disponibilizadas pelo software MATLAB® com as funções *fitnet* e *feedfowardnet,* A função *fitnet*, apresenta uma RN de ajuste, com uma função padrão para as camadas ocultas e uma função de transferência linear para as camadas de saída. Já o comando *feedfowardnet* apresenta uma função geral que implementa uma RN com duas ou mais camadas. Esta RN é aplicada na resolução de problemas de regressão e no reconhecimento e classificação de padrões.

Para o processamento dos dados foram utilizadas três camadas ocultas em cada RN, na qual há uma variação na quantidade de neurônios em cada camada, que serão entre 5 e 50 nas duas primeiras e entre 10 e 100 na terceira camada, em todos as variações utilizando o método de *Levenberg-Marquardt*. Essas variações possibilitam uma análise de sensibilidade entre as diferentes quantidades de neurônios, sendo expostas os resultados da arquitetura de neurônios que obtiver os menores erros.

A aplicação do método SVM é realizado através do comando *fitcecoc*, utilizando a matriz *input* e o vetor *output.* Este comando retorna um modelo completo, treinado e com os códigos de saída com os erros corrigidos. Os hiperparâmetros do método SVM são otimizados pela *toolbox* do MATLAB®. Para a identificação das classes do SVM, utilizam-se os resultados do conjunto, na função *predict,* que resulta em um vetor de rótulos de classe para os dados do modelo completo.

A aplicação do classificador RF foi realizada com a função *fitensemble*, empregando como método de agregação de conjuntos para dados com múltiplas classes e, também, com análise discriminante dos conjuntos e, como abordagem de 10 ciclos de aprendizagem.

4.2.3. Avaliação dos desempenhos dos métodos

Os desempenhos dos métodos são avaliados utilizando os indicadores *Root Mean Square Error* (RMSE) e Desempenho *Nash-Sutcliffe* (DNS). O RMSE (Equação 19) é um indicador consolidado em pesquisa que utilizam aprendizado de máquinas (BINDAL e SINGH, 2019). Sua unidade segue a dimensão dos dados observados e preditos, podendo ser entendido

como a média dos n erros calculados entre os dados observados (y) e os preditos ($\hat{y}$) ao quadrado, (WILLMOTT e MARSUURA, 2005).

$$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{n}} \qquad (19)$$

Já o DNS (Equação 20) compara o erro quadrado médio gerado no processo de aprendizagem de máquinas com a variação da sequência dos valores observados.

$$\mathrm{DNS} = 1 - \frac{\| y(t) - \hat{y}(t) \|^2}{\| y(t) - \overline{y(t)} \|^2} \qquad (20)$$

O coeficiente DNS possui valores entre -∞ e 1, sendo que quanto maior o valor, melhor é a concordância entre os dados observados e os preditos (SCHAEFLI e GUPTA, 2007; Nash e Sutcliffe, 1970). O DNS é usual em análises de resultados em estudos de modelagens, também já utilizado em pesquisas que utilizam RF (LIANG *et al.,* 2017) e o método *Support Vector Machine* (BRENTAN *et al.,* 2018).

4.3. Aplicação dos métodos de AM

Para a validação da proposta de localização de anomalias, a primeira aplicação da metodologia aqui apresentada fará uso de uma rede hidráulica de pequeno porte. Tal aplicação permite verificar e ajustar os conceitos, parâmetros e erros, a fim de que se obtenha maior confiabilidade dos resultados em redes mais complexas.

4.3.1. Descrição da Rede de distribuição de água

A metodologia é aplicada na rede fictícia apresentada por Al-Zahrani e Moied (2003) (Figura 6), composta por 15 nós, 23 tubos e 3 reservatórios, denominada de agora em diante de RZM. Optou-se por essa rede, visto que possui manobras hidráulicas de baixa complexidade e pela possibilidade de verificação dos resultados no processo de simulação, uma vez que também foi utilizada nas pesquisas de Cardoso *et al*. (2018) e Oliveira *et al*. (2018).

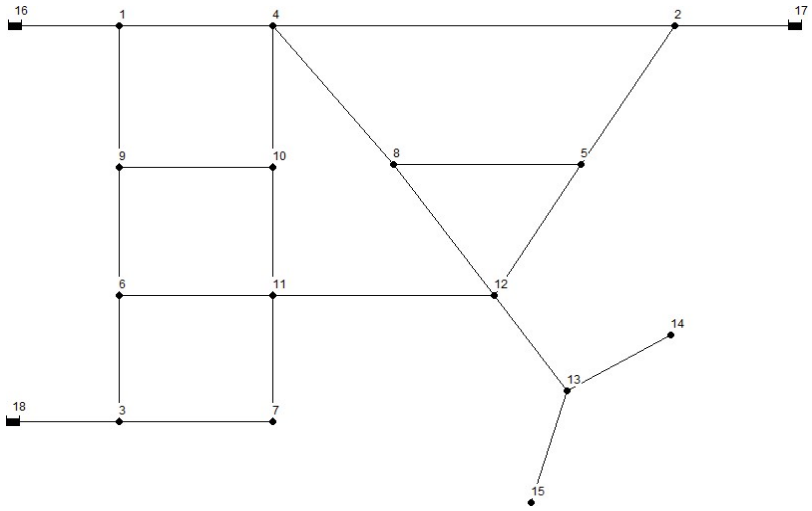

Figura 6 - Topologia da RDA utilizada. Fonte: (AL-ZAHRANI e MOIED, 2003).

Após a determinação dos nós monitorados com a aplicação do método de Cardoso *et al*. (2018) há a construção do arquivo (.msx) e identificação de contaminação química, proposto por Oliveira *et al*. (2018), assim possibilita a geração do banco de dados utilizado nas próximas etapas. O procedimento para alocação de sensores pode ser encontrado no Apêndice 1.

4.3.2. Construção do Código de Simulação

Após a seleção dos nós monitorados desenvolve-se o código para gerar o banco de dados e a aplicação dos métodos de aprendizado de máquinas. Primeiramente, há a inclusão dos dados dos nós monitorados e a definição dos dias de simulação. Optou-se pelo tempo de simulação de 15 dias, dada a necessidade da estabilização do comportamento padrão do cloro em toda a rede para que então sejam simuladas as situações anômalas. Para gerar essas situações anômalas, há a aplicação em diferentes concentrações do pesticida Parathion na metade da simulação, com duração de 12 horas.

Essas concentrações variam entre 0,1 mg/L a 19 mg/L, e são aplicadas individualmente em cada nó, com o objetivo de verificar a sensibilidade de variação de concentração do cloro. 10 cenários de contaminação para cada nó da rede RZM, esses cenários são analisados pelos sensores alocados, no qual se resulta em 150 cenários para cada sensor, em que contém todas as informações das interações químicas da água.

Ainda, realiza-se a modificação do arquivo de entrada (.msx). Essa etapa é necessária, visto que para o processo de alocação, utiliza-se uma concentração fixa do pesticida Parathion. Para isso, é necessário deixar limpa a sessão [SOURCES] do arquivo MSX, pois essas informações serão incluídas no código de simulação em ambiente MATLAB.

Um pseudo-algoritmo, com a aplicação dessas etapas no código de simulação, pode ser representado por:

Passo 1: *Abrir de arquivos .inp e .msx*

Passo 2: *Incluir os nós monitorados = '12' e '15'*

Passo 3: *Determinar o tempo de simulação = 15 dias*

Passo 4: *Indicar a quantidade de cenários:*

I. Números de cenários = 10
II. Nó fonte = cada nó da RDA repetidos o número de cenários
III. Concentração de contaminante = 0,1 a 19 mg/L
IV. Início da contaminação = após 168 horas
V. Tempo de injeção do contaminante = 12 horas
VI. Matriz = concatenar II, III, IV e V

For *i=1: I*

Passo 5: *Fonte e Concentração = Matriz (i , 1) , Matriz (i , 2)*

Passo 6: *Tempo de início e Duração = (Matriz (i ,3) : (Matriz (i , 3) +(Matriz (i , 4))) = 1*

Passo 7: *Definir o de padrão de injeção*

Passo 8: *Resolver a dinâmica de qualidade*

Passo 9: *Redefinir a de fonte de injeção*

End

Cada interação contaminante-cloro gera uma matriz com os componentes químicos utilizados nas equações estequiométricas do arquivo (.msx), essas equações estão apresentadas em Oliveira *et al*. (2018). Apresenta, ainda, a verificação das concentrações de cada componente químico, em cada nó monitorado, durante a simulação, ou seja, a simulação possui 2196 passos de tempo resultando em 2196 leituras dos sensores de monitoramento. Por fim, obtém-se 10 matrizes de interações de cada nó utilizado como fonte, e possui dimensões de 2 x 150, com as linhas representando a quantidade de sensores e as colunas com a quantidade total de cenários de contaminação em cada nó da rede.

4.3.3. Treinamento das RN

O método proposto para a identificação da fonte de contaminação através da variação da concentração de cloro consiste no treinamento de RNs alimentadas com os dados de

qualidade da água, obtidos através dos sensores de qualidade, que possuem as interações do cloro em situações que há a contaminação da RDA pelo pesticida Paration. Para tanto, são necessários os dados de entrada e saída (*input* e *output*), para o treinamento, validação e testes das RNs propostas. Assim, há a possibilidade de estudo dos erros e comparação dos resultados entre os métodos utilizados.

Com intuito de verificar a sensibilidade do método, tanto para os resultados quanto para os erros gerados, estabeleceu-se 3 camadas ocultas (*hidden layers*), no entanto testou-se diferentes quantidades de neurônios em cada camada. A quantidade de neurônios irá variar de 5 em 5 neurônios em cada camada.

O pseudo-algoritmo apresenta o modelo de codificação cujas funções para criação e treinamento das RNs, até então apresentadas, são expostas.

Passo 1: *Definir a quantidade de iterações para primeira camada (nmax_1).*

Passo 2*: Definir a quantidade de iterações para segunda camada (nmax_2).*

For *i=1,2,...,nmax_1*

For *j=1,2,..., nmax_2*

Passo 3: *RNA de ajuste = fitnet([(i * 5) , (j * 5) , (i + j * 5)], 'trainlm').*

Passo 4: *RNA de ajuste = train (RNA de ajuste, input, output).*

Passo 5: *RNA multicamadas = feedfowardnet([i * 5, j * 5, i + j * 5], 'trainlm').*

Passo 6: *RNA multicamadas = train (RNA multicamadas, input, output).*

End*j*

End*i*

Nos testes iniciais, houve 10 iterações para os Passos 1 e 2. Assim, as quantidades de neurônios, nas primeiras duas camadas, variam entre 5 e 50, e na terceira camada com a variação entre 10 e 100. Nos Passos 3 e 5 ocorre a criação das RNs, em que seguem as especificações de quantidade de camadas, quantidades de neurônios e o método de treinamento. Por fim, os Passos 4 e 6, treinam, validam e testam as Redes Neurais com os dados da matriz *input* e do vetor *output.*

Com essas iterações, é possível verificar erros apresentados para cada quantidade de neurônios utilizados. Assim, adota-se a arquitetura que obtiver o menor erro, finalizando com

a reaplicação do método, porém, com sua respectiva quantidade de neurônios. Deste modo, obtêm-se os resultados com a aplicação na arquitetura selecionada.

4.4. Resultados da aplicação dos métodos

A aplicação desta metodologia nas quatro abordagens de AM e na rede RZM, permite a comparação entre estas abordagens e posterior aplicação do método mais eficaz em RDA com maior grau de complexidade. Visto que a variação na quantidade de pesticida utilizado no ataque influenciava na concentração do cloro, utilizou-se, como dados de entrada, as concentrações de cloro no intervalo de reação com o Paration, mostrados entre as linhas tracejadas na Figura 7. Este intervalo conta com 82 passos de tempo para a rede RZM, resultando no mesmo valor de treinamentos e validações das abordagens.

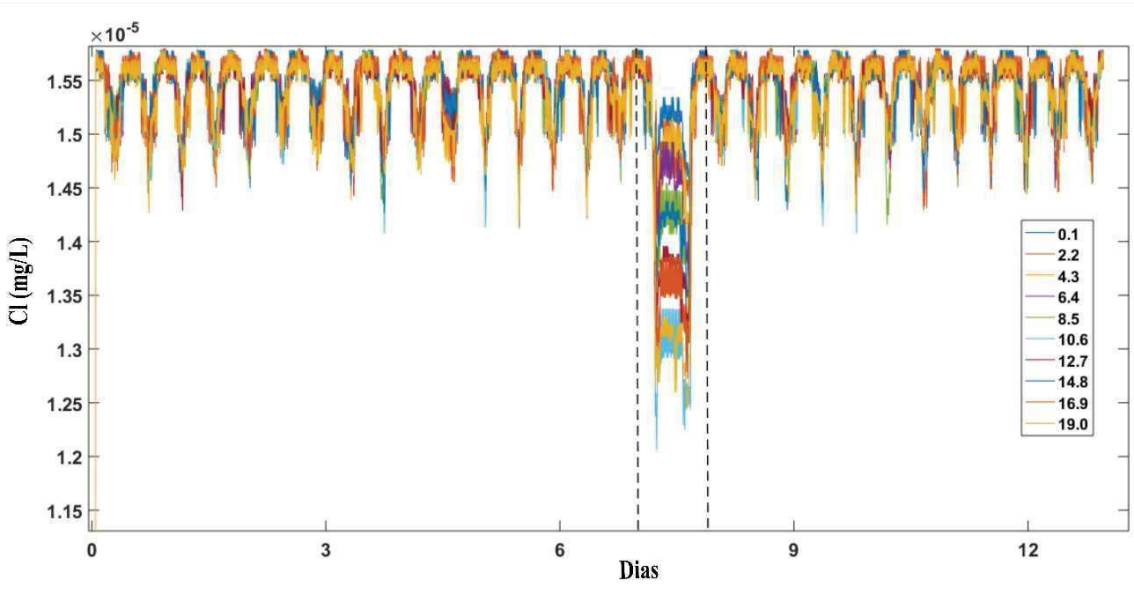

Figura 7 - Comportamento do Cloro ao longo de 15 dias de simulação.

A Figura 7 apresenta o comportamento de cloro do nó 2, onde ocorre a contaminação de 10 concentração de contaminantes e o monitoramento pelo sensor alocado no nó 12. O vetor *output* contém os dados das distâncias de cada nó entre o reservatório 16. A matriz *input* contém todo o comportamento do cloro monitorado pelos sensores nos nós 12 e 15. Os dados utilizados no processo de treinamento foram as concentrações de cloro com as intrusões de contaminante com 0.1, 2.2, 4.3, 14.8, 16.9 e 19.0 mg/L, resultando em 6 conjunto de dados. Já o processo de validação contou com as concentrações de cloro com as outras intrusões, resultando num conjunto de seis concentrações para o treinamento e quatro para validação.

Observar-se que os dados utilizados no treinamento são os 3 valores mínimos e os 3 máximos da concentração de contaminante aplicado. Este intervalo permite que os métodos de AM possam associar o comportamento do cloro, entre estes intervalos, com uma possível fonte de contaminação. Por este motivo, utiliza-se os valores de concentrações médios no processo de validação. Com isso, como essas concentrações foram aplicadas nos possíveis pontos de intrusão, ou seja, os 15 nós da rede, o processo de treinamento apresenta 90 cenários de contaminação e 60 cenários de contaminação no processo de validação dos métodos.

4.4.1. Resultados dos métodos de AM

O processo de variação na arquitetura de camadas e neurônios das RNs que apresentou menor valor de erro para a RN de ajuste, doravante denominada de FIT, possuía 10 neurônios na primeira camada, 40 na segunda e 18 na última. Já a arquitetura da RN do tipo *feedfoward,* daqui em diante denominada de FEED, resultou em 15, 50 e 25 neurônios por camada, respectivamente. A avaliação dos erros gerados no processo de alteração de neurônios por camada, pode ser encontrada no Apêndice A.2.

Os resultados dos métodos pelos indicadores de desempenho RMSE e DNS são expostos na Tabela 1. Os resultados do indicador RMSE foram divididos pela média dos valores reais de distâncias do reservatório, assim, quanto menor o valor mais próximo a abordagem está da média real. Já o indicador DNS, os resultados são entre 1 e -œ, ou seja, quanto mais próximo de 1, melhor o resultado do método.

Tabela 1 - Resultados dos indicadores de desempenho - RZM.

	RMSE Normalizado		DNS	
	Treinamento	Validação	Treinamento	Validação
SVM	0,21	0,24	0,70	0,62
FIT	0,13	0,16	0,86	0,81
FEED	0,13	0,16	0,86	0,83
RF	0,065	0,076	0,94	0,93

A Tabela 1 apresenta os valores de RMSE normalizado pela média dos valores de *output*. Observa-se que em ambos indicadores, tanto para treinamento quanto para validação, o método RF apresentou melhores resultados. As RNs apresentaram resultados próximos uma a

outra. Por outro lado, o SVM apresentou resultado inferior que os demais métodos. Contudo, buscou-se outra forma de avaliar o desempenho dos métodos, desta feita, utilizou os valores reais de distância, *output,* e os valores preditos pelos métodos de AM, considerando predição correta quando a raiz quadrada da diferença ao quadrado entre estes valores seja no máximo de uma unidade de distância, podendo ser observado na Figura 7.

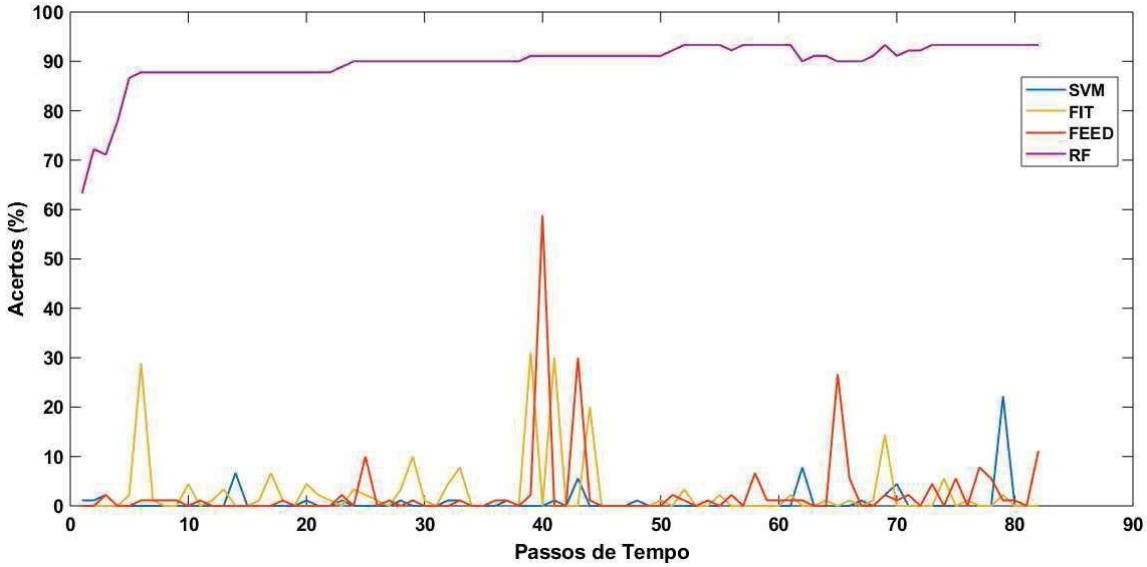

Figura 8 - Probabilidade de detecção RZM - Treinamento.

Nesta forma de avaliação de desempenho, o classificador RF se mostrou muito superior aos demais métodos, apresentando aumento constante à medida que mais passos de tempo são incluídos no processo de treinamento. Por outro lado, as RNs e o método SVM apresentam apenas picos de identificação das fontes de contaminação. Esse comportamento ocorre também no processo de validação (Figura 9), em que também se pode observar as porcentagens de identificação do RF e os picos das RNs e o SVM.

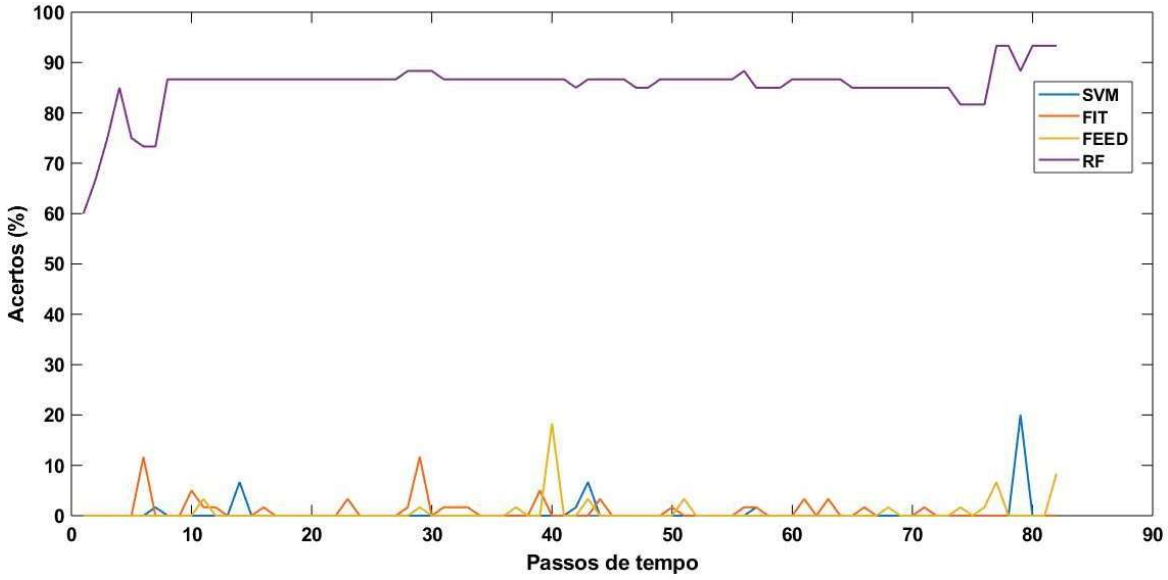

Figura 9 - Probabilidade de detecção RZM - Validação.

Quando calculado a raiz da diferença ao quadrado entre valores preditos e os observados, individualmente, com até uma unidade de distância, sua porcentagem de acertos ultrapassa 70%, tanto no processo de treinamento quanto no de validação logo nos primeiros passos de tempo analisados. Já as RNs possuem um pico de acertos, aproximadamente, no passo de tempo 40 no processo de treinamento, sendo 58% de acertos pela rede FEED e 32% para a rede FIT, e alcançam uma máxima de 20% no processo de validação. Por fim, o método SVM apresenta 20% de acertos, tanto no treinamento quanto na validação, próximo ao passo de tempo 80.

4.4.2. Probabilidade de detecção

Procurando analisar a melhora do processo de localização à medida que mais passos de tempo são inclusos, considerou-se os acertos de localização com até 1 metro da distância real para cada passo de tempo aplicados nos métodos de AM. Essa análise considera a quantidade de passos de tempo em que os nós são localizados. Na Figura 10, pode-se observar esta quantidade de nós que cada método localizou, com relação às quantidades de passos de tempo em que ocorrem estes acertos.

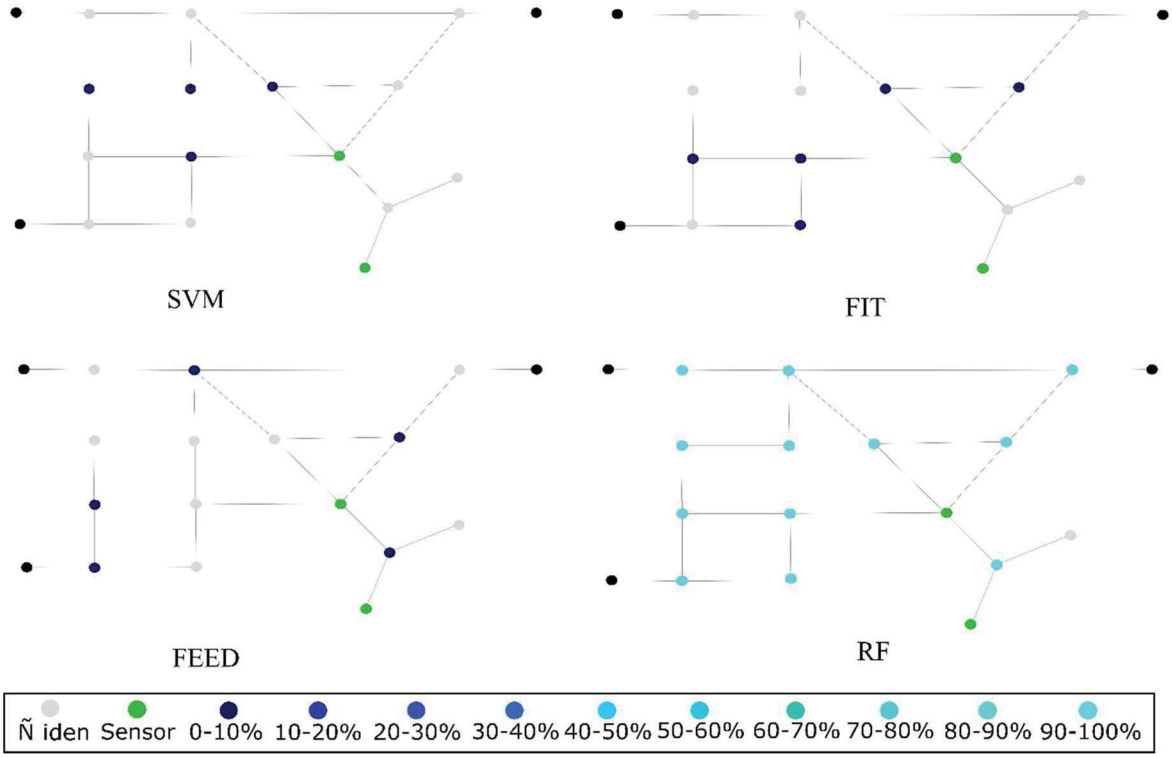

Figura 10 - Nós reconhecidos por porcentagem de passos de tempo.

O método SVM (Figura 10 (a)) localizou 4 fontes de contaminação em até 10% dos passos de tempo. As redes FEED e FIT ((Figura 10 (b) e (c), respectivamente) localizaram 5 fontes de contaminação, também em até 10% dos passos de tempo. Resultados já esperados, uma vez que estes desempenhos também puderam ser visualizados quando expostos as porcentagens de acertos. Em contrapartida, o classificador RF localizou 14 dos 15 nós fontes de contaminação entre 90%-100% dos passos de tempo utilizados no processo de validação. Este resultado é possível, visto que os nós sensores recebem contribuição de água dos reservatórios com até 8 passos de tempo após o início da simulação. Outra observação com respeito o nó 14, não identificado em nenhuma abordagem, visto ser nó de extremidade e não receber fluxo de água.

4.5. Análise de espalhamento

A localização acurada da fonte de contaminação em menor tempo possível é o fator mais influente nas pesquisas relacionadas. Pois, mesmo com a alocação otimizadas dos sensores e a efetiva funcionalidade dos métodos de localização de fonte, há um espaço de tempo até que os nós monitores recebam contribuição dos nós que foram contaminados. Procurando entender esse comportamento, realizou-se uma análise temporal do percurso do contaminante até a identificação e localização da fonte.

Para analisar o espalhamento do contaminante na RZM os dados de contaminação do reservatório 16 e os nós 3 e 13 foram selecionados. A escolha destes nós ocorreu pelo intuito de observar a influência que cada um possui quando utilizados como fonte de contaminação. Optou-se por escolher um dos reservatórios, devido sua influência na contribuição e por ser um ponto provável em caso de contaminação intencional. Devido às características da rede, como comprimento dos tubos, vazão e demanda, o contaminante a percorre em apenas um passo de tempo para todos os nós observados. A Figura 11 exibe o este espalhamento para cada nó selecionado.

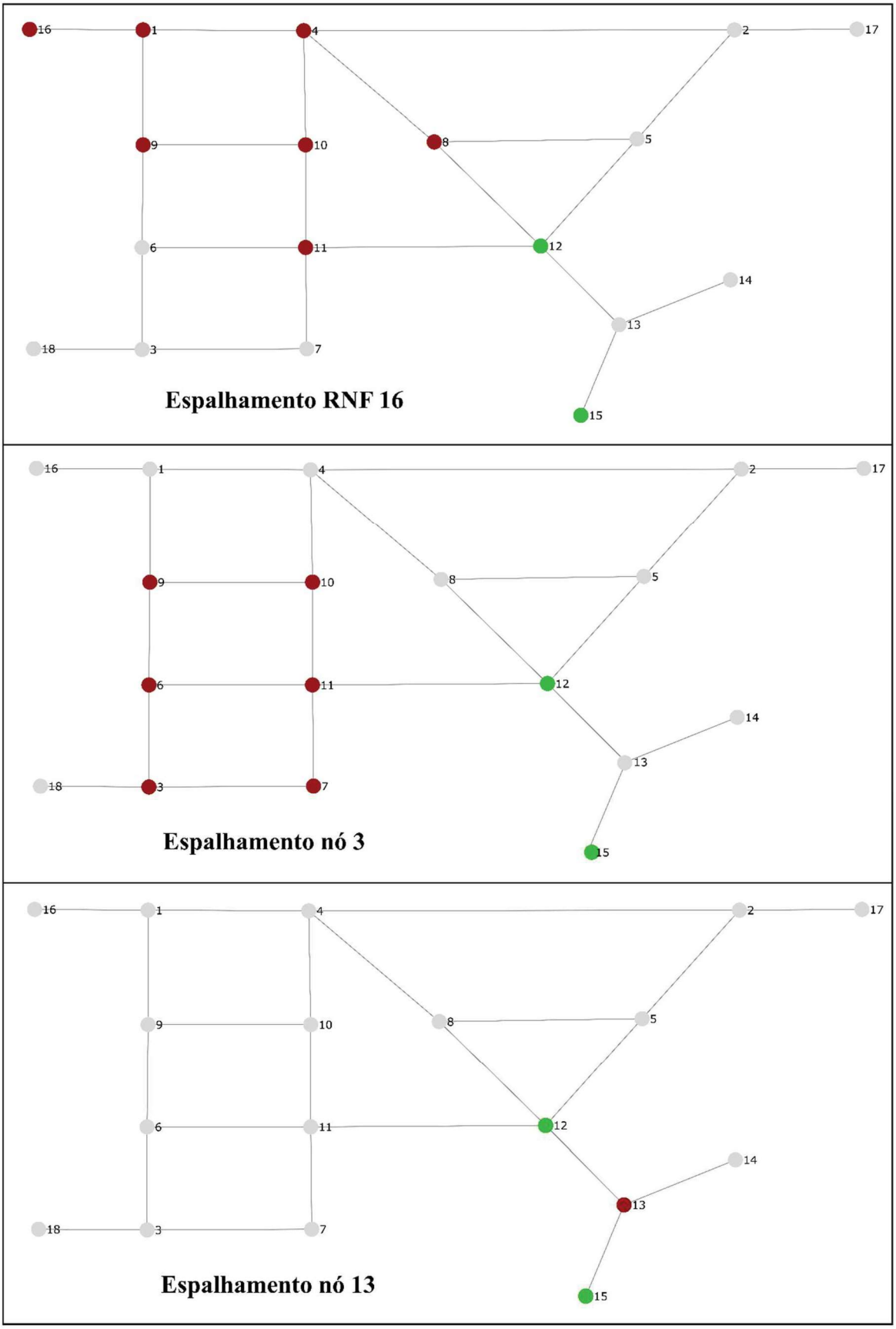

Figura 11 - Espalhamento - RZM.

A Figura 11 expõe quantos nós recebem contaminante até que os nós monitores sejam atingidos. Como esperado, o espalhamento do contaminante quando utilizado o reservatório 16 é maior que os demais nós analisados. A contaminação do nó 3 também possui um grande espalhamento, principalmente pela sua proximidade com o reservatório. Por outro lado, caso contaminado o nó 13, apenas o nó monitor 15 receberá contribuição uma vez que o nó 14 não recebe fluxo de água.

4.6. Discussões parciais

A metodologia proposta foi aplicada, primeiramente, na rede RZM de baixa complexidade para ajuste dos parâmetros, dos métodos e na determinação dos procedimentos que são utilizados em redes maiores e mais complexas. Para isso, algumas discussões devem ser expostas, em relação a abordagem de *output* até então utilizada e, também, em relação aos métodos de aprendizado de máquinas.

4.6.1. Abordagens de *Output*

Um aspecto importante desse trabalho, discutido aqui com maior profundidade, é a forma de representar o *output* para o aprendizado de máquinas. Normalmente, a localização de anomalias é feita pela rotulação dos sinais de entrada, pelo índice do nó fonte da anomalia ou com um valor dimensional. No entanto, pode haver interpretações errôneas com essas abordagens, principalmente na aplicação de aprendizado de máquinas, uma vez que a indexação dos nós não segue, necessariamente, um critério de proximidade. Dessa forma, nós distantes, mas com índices sequenciais podem ser interpretados pelos métodos de aprendizado de máquinas como próximos.

Ainda que a abordagem desse trabalho se dê pela utilização da distância euclidiana entre os nós e um ponto de referência, tal distância delimita um raio, ou seja, outros nós que estejam neste mesmo raio terão a mesma distância entre o ponto de referência, o que ainda pode levar a erros consideráveis nos métodos de aprendizado de máquinas.

A Figura 12 pode representar a situação dita para as abordagens, uma vez que seja contaminado o nó 5, e os métodos de AM apontado o nó 7 como fonte de contaminação, utilizado as identificações dos nós no vetor *output*, o erro seria descrito pela equação 18:

$$\text{erro} = ||5 - 7|| = 2 \qquad (18)$$

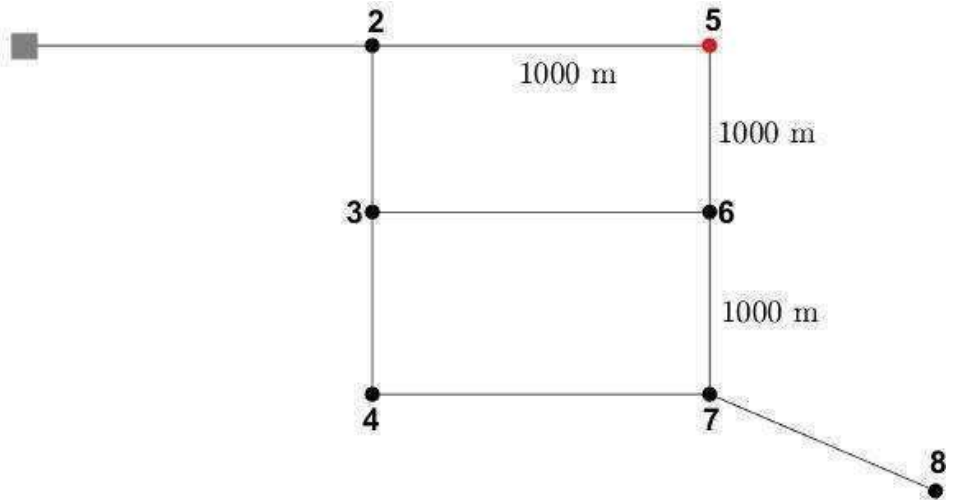

Figura 12 - Rede Explicativa.

Porém, estes nós estão distantes um do outro. Caso o nó 2 fosse identificado como a fonte de contaminação, esse erro passaria a ser 3, mesmo que este esteja mais próximo do nó 5. Essa situação, quando utilizada as distâncias lineares, em que contaminou-se o nó 5 e a metodologia indicou como fonte o nó 7, ou seja, as abordagens de AM indicaram a fonte de contaminação a 2000 metros, o erro passaria para (equação 19):

$$\text{erro} = ||0 — 2000|| = 2000 \qquad (19)$$

Porém, se as abordagens indicassem como fonte de contaminação um nó a 1000 metros de distância do nó 5, os nós 2 e 6 poderiam ser selecionados como fontes. Há assim uma redução no erro apresentado, mas duas opções de fontes de contaminação.

Diante destas situações, são utilizadas novas abordagens de *output* com o intuído de diminuir os erros gerados pelas situações apresentadas. Foram realizadas novas abordagens utilizando coordenadas geográficas e as distâncias, principalmente, em conjunto, para garantir maior confiabilidade dos resultados.

5. APLICAÇÃO DE RF PARA LOCALIZAR FONTES DE CONTAMINAÇÃO

O processo de validação da metodologia proposta utilizando a rede RZM, mostrou a eficácia do classificador RF. Por isso, este classificador será empregado na localização de fontes de contaminação em outras RDA, maiores e mais complexas. São utilizadas novas abordagens de *output* e uma nova proposta na aplicação do classificador RF.

5.1. Rede de Distribuição de Água – Jiuyou

A metodologia proposta foi aplicada na rede Jiayou (JYN), apresentada por He *et al.* (2018). Esta rede (Figura 13) possui 300 nós, 451 tubos e dois reservatórios, no qual se considerou todos os nós e os reservatórios como possíveis pontos para a alocação de sensores e fontes de contaminação, resultando em 302 cenários de contaminação. A metodologia de Cardoso *et al.* (2018) também foi utilizada para a determinação dos nós monitores A quantidade de sensores alocados seguiu a pesquisa de He *et al.* (2018), onde alocam 6 sensores de monitoramento, com diferentes fatores para decisão da alocação.

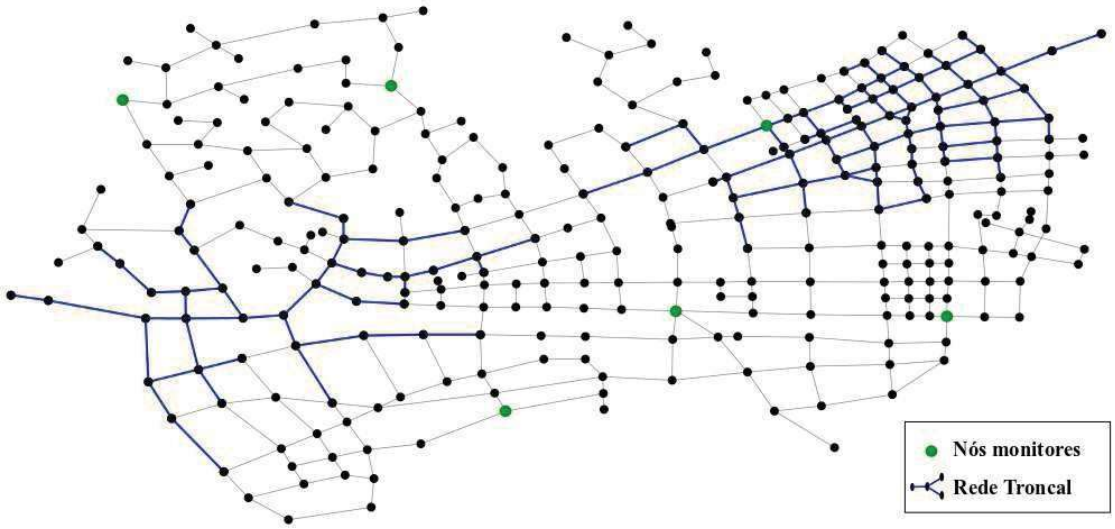

Figura 13 - RDA Jiuyou (He *et al.,* 2018).

Os nós selecionados (Figura 13) para a alocação de sensores foram 53, 30, N7, N272, N266 e N44. Estes nós se mostraram coerentes com a pesquisa de He *et al.* (2018), principalmente os nós N7 e N266, que também foram selecionados para o monitoramento. Esse conjunto de sensores possui uma taxa de cobertura de 86% dos nós da rede. A Figura 13 mostra

ainda as tubulações da rede que possuem maiores vazões, distribuindo água para todos os outros pontos, denominada Rede Troncal.

5.1.1. Simulação de qualidade de água

Nas simulações foram empregadas três concentrações de Paration em todos os nós e nos reservatórios. As concentrações são 0.1, 12.4 e 19.0 mg/L, nas quais permitem a análise da sensibilidade do classificador RF. Esta quantidade foi estipulada por representar um intervalo que considera a concentração de solubilidade e dose letal. A duração da simulação e o tempo de início da contaminação é padrão para todos os cenários, sendo 24 horas de simulação, a aplicação do Paration iniciando às 6 horas após o início da simulação e a intrusão ocorreu durante 12 horas. Estes parâmetros permitem a completa análise do comportamento, espalhamento e interações do contaminante na rede.

Os sensores de qualidade coletam a concentração de cloro em intervalos de 10 minutos, aqui chamados de passos de tempo. Cada simulação é composta por 144 passos de tempo durante a simulação. No entanto, só serão utilizadas as informações a partir do início da contaminação, ou seja, como a contaminação ocorre 6 horas após o início da simulação, ao final da simulação se obtém um banco de dados, em que cada cenário de contaminação possui 111 leituras das concentrações de cloro para cada sensor alocado. A exclusão das primeiras 6 horas de simulação ocorre, primeiramente, devido à concentração de cloro se estabilizar algumas horas após o início da simulação, para esta estabilização é considerada a concentração contínua de cloro. No entanto, a identificação de anomalias nos dados ocorre quando se encontram divergências que podem ser associadas a episódios, neste caso, de contaminações, corroborando com a retirada dos dados e utilizando os que apresentarem anomalias.

5.1.2. Abordagem RF, *Input* e *Outputs*

O processo de treinamento e validação do RF com o banco de dados gerados com os cenários de contaminação na rede JYN, ocorre de acordo a proposta de localização em tempo real utilizada na rede RZM, ou seja, de acordo os passos de tempo que houveram monitoramento da concentração de cloro pelos sensores de monitoramento. Por isso, na rede JYN, ocorrem 111 treinamentos e validações, em que esse processo de treinamento inicial conta como dados de entrada apenas as concentrações de cloro relativas ao primeiro passo de tempo e o último processo com as 111 concentrações concatenadas.

Neste novo processo serão utilizadas três abordagens de *Output*, segue-se utilizando a metodologia proposta por Barros *et al*. (2019), com a utilização da distância das fontes de contaminação a um ponto específico, utiliza-se, também, a distância até um reservatório, esta abordagem será doravante é chamada MinMax. As outras abordagens de *output* utilizam as coordenadas cartesianas desses nós, sendo um processo de treinamento e validação para a coordenada x (CoordX) e outro processo para a coordenada y (CoordY).

O software EPANET 2.0 não utiliza as coordenadas nos cálculos hidráulicos e qualidade. Por isso, os arquivos das redes podem ou não conter as coordenadas reais dos componentes da WDN, mesmo assim, caso essas informações não sejam oferecidas, há a estipulação dessas coordenadas pelo software em um plano cartesiano (ROSSMAN, 2000). As abordagens, CoordX e CoordY, que utilizam estas coordenadas no processo de treinamento, contarão com os valores das coordenadas, uma vez que estes dados são disponibilizados pelo arquivo da rede.

O método RF terá, em seu primeiro treinamento e validação, a quantidade fixa de ciclos de aprendizagem, utilizando a abordagem de *output* MinMax, e um segundo processo com a quantidade de ciclos aumentando a cada passo de tempo que se inclua no treinamento e validação, também com a abordagem MinMax, esta abordagem será identificada como ICcl Após estes primeiros treinamentos, será realizado uma análise de desempenho, a fim de priorizar a utilização do método de ciclos que apresentar menor erro nas abordagens CoordX e CoordY. Os métodos utilizados nos processos de treinamento e validação continuam com a agregação de conjuntos para problemas com múltiplas classes, com análise descriminante, porém, a abordagem ICcl conta com 11 ciclos de aprendizagem no primeiro processo de treinamento e validação e posterior acréscimo de ciclos a cada passo de tempo incluído no treinamento, sendo o último com 121 ciclos.

5.1.3. Avaliação de desempenho

O desempenho das abordagens pelos indicadores RMSE e DNS (Tabela 2) indica uma mínima diferença entre a aplicação da abordagem de MinMax e ICcl, mesmo assim a variação no ciclo de aprendizagem do RF foi utilizado do processo de treinamento e validação com as abordagens de *output* CoordX e CoordY.

Tabela 2 - Resultados dos indicadores de desempenho - JYN.

	RMSE normalizado		DNS	
	Treinamento	Validação	Treinamento	Validação
MinMax	0,29	0,36	0,53	0,33
ICcl	0,27	0,36	0,56	0,36
CoordX	0,36	0,43	0,53	0,35
CoordY	0,35	0,40	0,22	0,001

Como na aplicação da metodologia na rede RZM, os resultados do indicador RMSE foram divididos pela média dos valores reais de distâncias e coordenadas, deste modo, valores menores representam maior proximidade da média real. A abordagem ICcl apresentou menores valores se considerado os indicadores RMSE e DNS.

A avaliação através da probabilidade de detecção, quando se considera a raiz da diferença ao quadrado entre o valor observado e o predito, considerada como acerto até 5 metros para a aplicação na rede JYN. Este aumento de valor em relação a abordagem da rede RZM, acorre devido a maior dimensão da rede JYN e por não haver nós próximos uns dos outros com distâncias menores que 5 unidades de medida. A Figura 14 mostra a porcentagem de cobertura do treinamento das abordagens a cada passo de tempo, na qual se pode observar o aumento das porcentagens à medida que mais passos de tempo são utilizados no treinamento do RF.

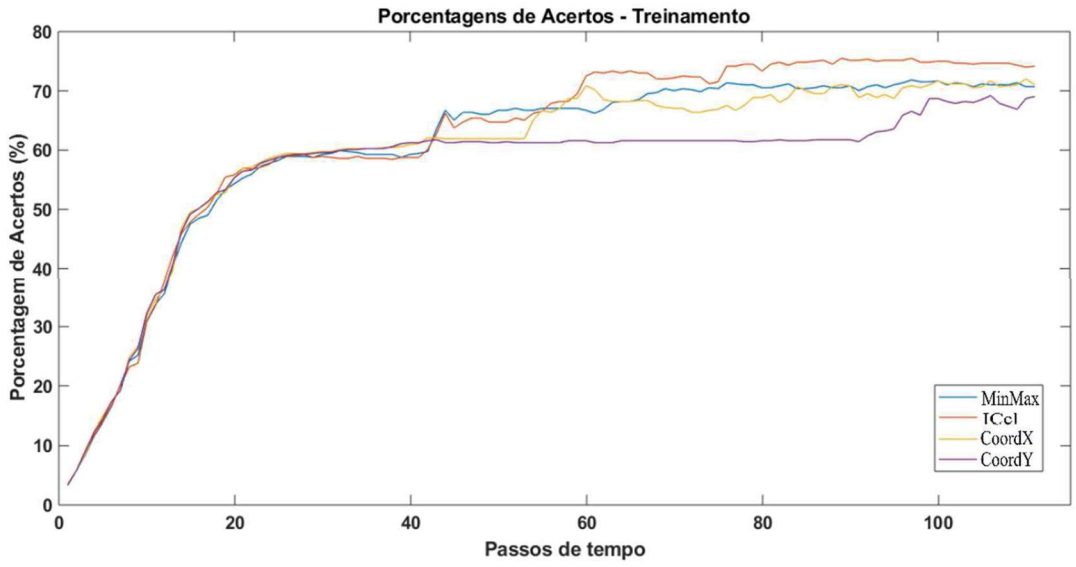

Figura 14 - Probabilidade de detecção JYN – Treinamento.

Na validação do RF (Figura 15), as porcentagens de acertos seguem um padrão para todas as abordagens, porém alcançam a porcentagem máxima no passo de tempo 19, seguindo de um decréscimo à medida que mais passos de tempo são incluídos no processo de validação. Esse comportamento pode ser associado a quantidade de dados utilizados no processo de validação, que continham apenas as informações de uma concentração do pesticida. Esta única concentração pode causar redundância nos dados ao longo da simulação e interferir no processamento do RF. Pode-se, ainda, observar que as abordagens MinMax e ICcl possuem comportamento similares.

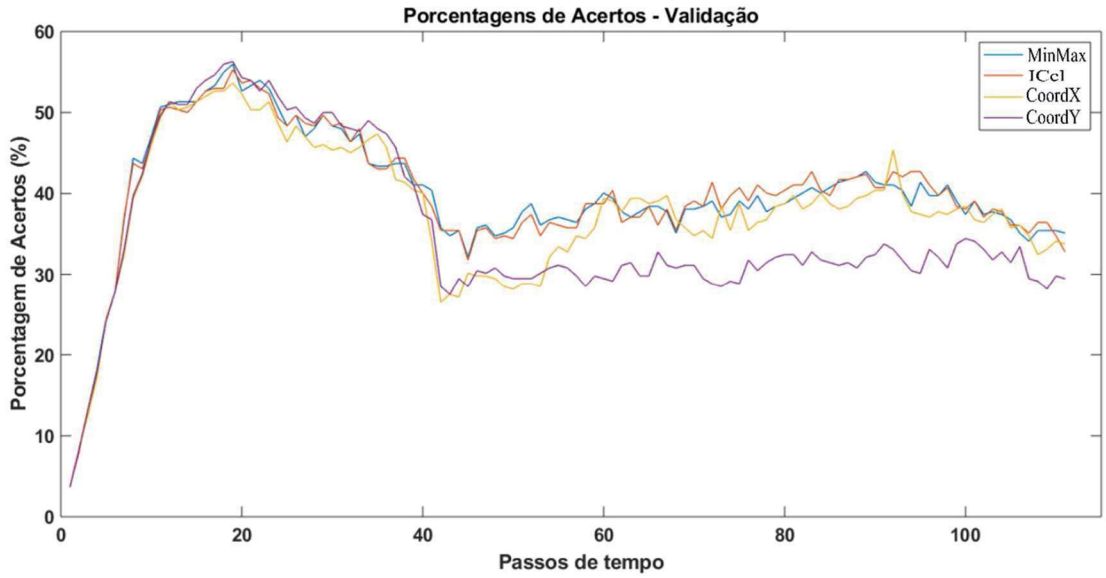

Figura 15 - Probabilidade de detecção JYN – Validação.

Observa-se que a abordagem ICcl possui comportamento semelhante a abordagem MinMax, tanto para treinamento quanto para validação, porém a abordagem ICcl apresenta melhor desempenho que as demais abordagens. Esta semelhança também pode ser observada nos resultados dos indicadores RMSE e DNS. Por outro lado, a abordagem CoordY que apresentou melhores resultados nos indicadores RMSE e DNS, apresenta um desempenho inferior às outras abordagens a partir do passo de tempo 50.

5.1.4. Probabilidade de detecção

Observando individualmente os passos de tempo para cada abordagem, como realizado na validação da metodologia, é possível analisar as porcentagens de passos de tempo em que os nós são identificados no processo de validação do classificador RF, isto para a rede JYN. Na abordagem MinMax (Figura 16 (a)), os nós que são localizados na grande maioria dos passos

tempo se encontram na região centro-direita ou próximos aos nós monitores. Nesta abordagem, 83,1% dos cenários de contaminação são localizados, sendo que 17,2% são localizados entre 90% e 100% dos passos de tempo.

A abordagem ICcl (Figura 16 (b)), como esperado, possui a localização por passos de tempo similar ao da abordagem MinMax, com exceção de alguns nós que passam a ser identificados em mais passos de tempo. Outra semelhança com a abordagem anterior é a porcentagem de cenários de contaminação localizados, com 83,8 %, sendo 16,2% localizados entre 90 % a 100% dos passos de tempo.

A Figura 16 (c) expõe as localização das fontes de contaminação por porcentagem de passos de tempo para a abordagem CoordX, onde 81,8% dos nós da rede são identificados como fontes de contaminação, um resultado menor que o MinMax e ICcl, porém tem uma leve superioridade nos cenários localizados entre 90% a 100% dos passos de tempo, sendo 17,5% do total de nós contaminados. A abordagem CoordY (Figura 16 (d)), identifica 76,83% dos cenários de contaminação, uma redução significativa em relação às outras abordagens, mas localizou 16,9\% dos cenários de contaminação entre 90 % e 100 % dos passos de tempo.

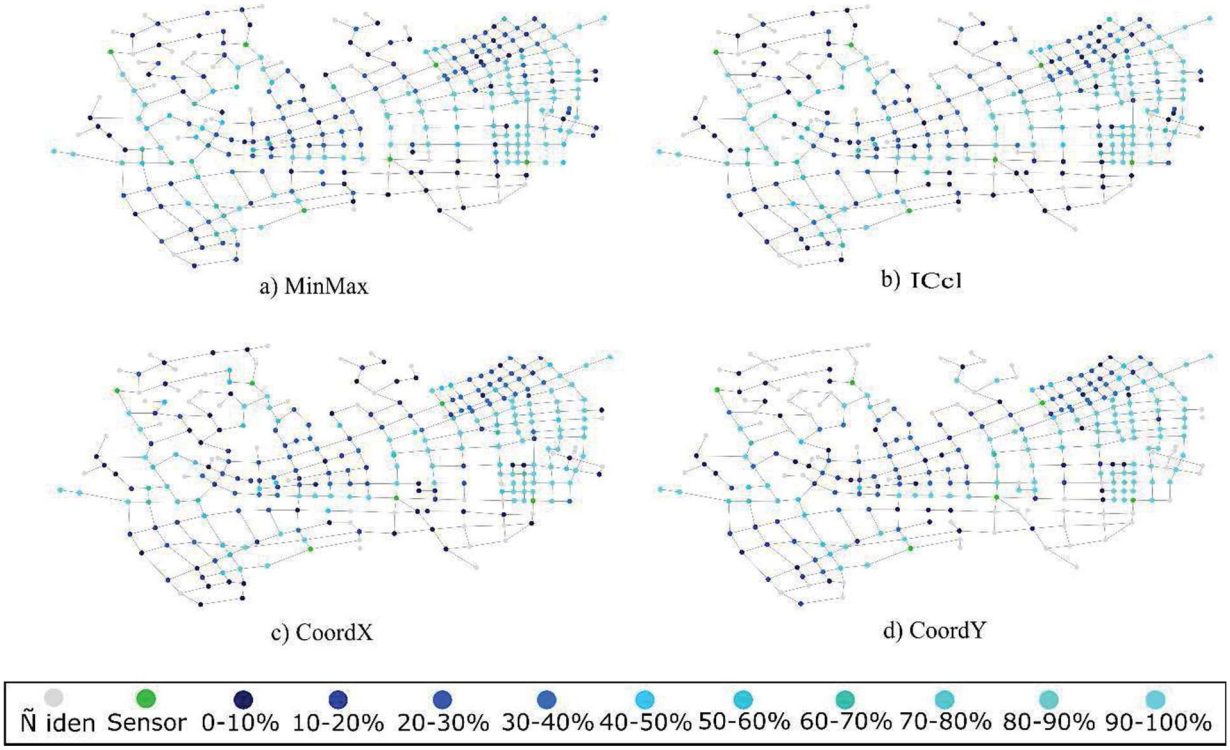

Figura 16 - Localização dos nós por passo de tempo - JYN.

Todas as abordagens apresentaram resultados coerentes, uma vez que nenhum nó de extremidade teve a contaminação identificada. Outro ponto a observar, são as contaminações nos reservatórios, ambas identificadas em no mínimo 90% dos passos de tempo, visto que água dos reservatórios chegam em até 8 passos de tempo nos nós monitores.

5.1.5. Impacto de contaminação

Procurando investigar o espalhamento do Paration na RDA, alguns nós foram selecionados para a observação do comportamento do contaminante até sua identificação pelos sensores ou quando não são identificados. Estes nós, dividem-se entre o reservatório RNF 1 e o nó N117, que são identificados 8 passos de tempo após o início da contaminação, e os nós N246, N14, N211 e N75 que não são identificados como fontes de contaminação em nenhuma das abordagens testadas.

A Figura 17 evidencia o espalhamento da contaminação pelos passos de tempo após o início da aplicação nos referidos nós. Pode se observar o grau de influência que uma contaminação no RNF 1 pode causar na rede JYN, esse comportamento é esperado, visto que o reservatório é a fonte de água para os nós da rede. O tamanho do espelhamento ainda pode

ser associado as tubulações pertencentes a Rede Troncal, em que parte se iniciam na ligação do reservatório aos nós da rede. Porém, até sua determinação como fonte de contaminante passam-se 7 passos de tempo, afetando 65 nós até a identificação de contaminante e posterior localização da fonte.

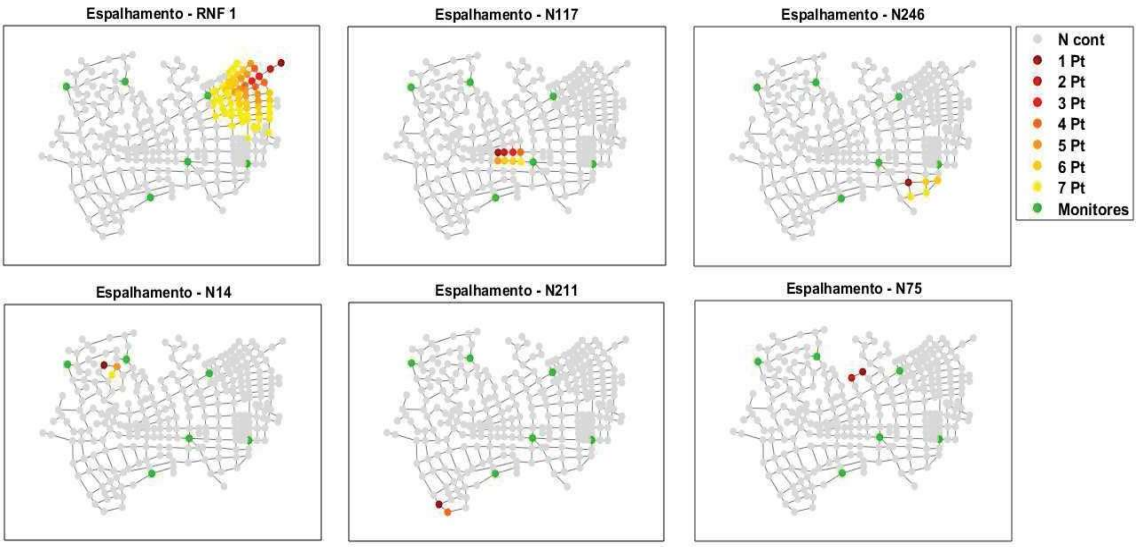

Figura 17 - Espalhamento de Contaminante - JYN.

Pode-se observar, ainda, que mesmo com a proximidade do nó N117 ao nó monitorado N272, 7 passos de tempo decorrem até sua localização como fonte, contaminando outros 8 nós. Os nós N246, N14, N211 e N75 contaminam poucos nós, principalmente se comparado ao RNF 1 e ao nó N117, porém estas contaminações não são identificadas pelos sensores, impedindo a localização da fonte e que as medidas de segurança possam ser tomadas.

6. CONCLUSÃO

Esta pesquisa apresenta um modelo computacional para localização de fontes de contaminação em caso de intrusão de compostos químicos, que se mostrou eficaz na determinação das fontes. Para alcançar esses resultados algumas etapas tiveram que ser realizadas, iniciando-se no processo de simulação e obtenção dos dados de qualidade. Nesta etapa, as informações de qualidade são obtidas pelo monitoramento através de sensores alocados nos nós da rede. Porém, há redes de distribuição que não possuem esses sensores alocados, sendo necessário utilizar uma metodologia para alocação otimizada de sensores.

A segunda etapa para a elaboração do modelo é a análise dos dados emitidos pelos sensores. Nesta pesquisa utilizou-se três métodos de aprendizado de máquinas para identificar padrões nos dados de cloro e associá-los às fontes de contaminação simuladas. Os resultados dos indicadores RMSE e DNS apresentados pelo algoritmo RF foram mais precisos que as RN e o SVM. Essa comprovação ocorreu também mediante uma abordagem de desempenho utilizada, na qual considerou-se assertiva quando os métodos de aprendizado de máquinas determinavam a localização do nó contaminado em um intervalo aceitável.

A abordagem ICcl apresentou melhor desempenho que as demais abordagens. Esta abordagem utiliza distâncias entre os sensores e o local da intrusão com aumento no ciclo de aprendizado do algoritmo RF. Os sensores alocados na rede JYN obtinha uma taxa de cobertura de 86% dos nós, e a abordagem ICcl localizou 76% dos cenários de contaminação. Esses resultados representam que dos 260 nós cobertos pelos sensores, apenas 30 não são localizados pelo método. As abordagens MinMax e CoordX localizaram cerca de 71% dos nós e a abordagem CoordY localizou 69,3%.

A última etapa deste modelo buscou analisar os impactos causados por uma contaminação até que sua fonte seja identificada e localizada. Nesta análise, alguns nós contaminados foram selecionados a fim de demonstrar os percursos e a quantidade de nós são contaminados por passos de tempo até que os nós sensores recebam contribuição do contaminante e seja possível a localização da fonte através dos métodos de aprendizado de máquinas.

Mesmo com a localização ocorrendo de forma precisa para os cenários de contaminação simulados, a quantidade de sensores na RDA é o fator de maior influência na localização rápida das fontes. Pois, devido ao número limitado de sensores, há um grande intervalo de tempo até que os sensores identifiquem a contaminação, ocorrendo também nós cuja a contaminação

nunca são reconhecidas. Essa problemática poderá sempre existir devido a impossibilidade de monitoramento de todos os nós da rede, porém pesquisas com o intuito de alocação otimizada de sensores vem crescendo cotidianamente e alcançando, cada vez mais, ótimos resultados.

O modelo proposto pode contribuir não só para a localização de fontes quando houver intrusão de composto químicos, mas também para qualquer componente que afete os parâmetros de potabilidade da água. Componentes biológicos podem afetar a concentração de cloro, turbidez e oxigênio presentes na água e, assim, é possível associar as variações de concentração a possíveis fontes. Sensores biológicos também podem ser utilizados, até em conjunto com os sensores químicos, aumentando ainda mais a segurança da RDA. A utilização de outros métodos de aprendizado de máquinas também pode contribuir na determinação mais assertiva das fontes, principalmente devido aos avanços que esses métodos vêm alcançando a cada dia.

REFERÊNCIAS

ABDULSHAHEED, A.; MUSTAPHA, F.; GHAVAMIAN, A. A pressure-based method for monitoring leaks in a pipe distribution system: A Review. **Renewable and Sustainable Energy Reviews**, v. 69, p. 902-911, 2017.

ABRAHAM, Ajith. Artificial neural networks. **handbook of measuring system design**, 2005.

AGRESTA, A., FATTORUSO, G., LANZA, B., FABBRICINO, M., TRIFUOGGI, M., DE VITO, S., &DI FRANCIA, G. Applying numerical models and optimized sensor networks for drinking water quality control. **Procedia Engineering**, *119*, 918-926, 2015.

AISOPOU, Angeliki; STOIANOV, Ivan; GRAHAM, Nigel JD. In-pipe water quality monitoring in water supply systems under steady and unsteady state flow conditions: A quantitative assessment. **Water research**, v. 46, n. 1, p. 235-246, 2012.

AKO, Andrew Ako; NKENG, George Elambo; TAKEM, Gloria Eyong Eneke. Water quality and occurrence of water-borne diseases in the Douala 4th District, Cameroon. **Water science and technology**, v. 59, n. 12, p. 2321-2329, 2009.

AL-ALAWI, Ali; AL-ALAWI, Saleh M.; ISLAM, Syed M. Predictive control of an integrated PV-diesel water and power supply system using an artificial neural network. **Renewable energy**, v. 32, n. 8, p. 1426-1439, 2007.

AL-ANI, Dhafar; HABIBI, Saeid. Optimal pump operation for water distribution systems using a new multi-agent Particle Swarm Optimization technique with EPANET. In: **Electrical & Computer Engineering (CCECE), 2012 25th IEEE Canadian Conference on**. IEEE, 2012. p. 1-6.

ALLMANN, Ytimothy P.; CARLSON, Kenneth H. Expanded summary: Modeling intentional distribution system contamination and detection. **Journal-American Water Works Association**, v. 97, n. 1, p. 58-62, 2005.

AL-MUKHTAR, Mustafa. Random forest, support vector machine, and neural networks to modelling suspended sediment in Tigris River-Baghdad. **Environmental monitoring and assessment**, v. 191, n. 11, p. 673, 2019.

AL-ZAHRANI, M. A; MOIED, K. Optimizing water quality monitoring stations using genetic algorithms**. The Arabian Journal for Science and Engineering**, v. 28, n. 1, p. 57–75, 2003.

ALY, Anne; GREEN, Lelia. Fear, anxiety and the state of terror. **Studies in Conflict & Terrorism**, v. 33, n. 3, p. 268-281, 2010.

ANTONY, R., NANDAGOPAL, M. G., SREEKUMAR, N., & SELVARAJU, N. (2014). Detection principles and development of microfluidic sensors in the last decade. **Microsystem Technologies**, *20*(6), 1051-1061.

ARAL, Mustafa M.; GUAN, Jiabao; MASLIA, Morris L. Optimal design of sensor placement in water distribution networks. **Journal of Water Resources Planning and Management**, v. 136, n. 1, p. 5-18, 2009.

ARAUJO, L. S.; RAMOS, H.; COELHO, S. T. Pressure control for leakage minimisation in water distribution systems management. **Water resources management**, v. 20, n. 1, p. 133-149, 2006.

ASSANTE, Michael J. Infrastructure protection in the ancient world. In: **System Sciences, 2009. HICSS'09. 42nd Hawaii International Conference on**. IEEE, 2009. p. 1-10.

BAJPAI, Saumya; JAIN, Kreeti; JAIN, Neeti. Artificial neural networks. **International Journal of Soft Computing and Engineering (IJSCE)**, v. 1, n. NCAI2011, 2011.

BARROS, D. B., CARDOSO, S. M., OLIVEIRA, E., BRENTAN, B., LUVIZOTTO, E., RIBEIRO, L. Localização de fonte de contaminação em rede de distribuição de Água utilizando aprendizagem estatística. **XVI Seminário Ibero-Americano sobre Sistemas de Abastecimento e Drenagem.** 2019.

BELGIU, Mariana; DRĂGUŢ, Lucian. Random forest in remote sensing: A review of applications and future directions. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 114, p. 24-31, 2016.

BIEHL, Luciano Volcanoglo. **A ciência ontem, hoje e sempre**. Editora da ULBRA, 2003.

BINITHA, S., & SATHYA, S. S. A survey of bio inspired optimization algorithms. **International Journal of Soft Computing and Engineering**, *2*(2), 137-151, 2012.

Bonfim, C. E. S. Simulação de CFD para análise de ambiente contaminado com agentes de guerra química: influência dos perfis de escoamento e determinação de áreas de risco sobre um sistema de distribuição de água. Instituto militar de engenharia. Rio de Janeiro. 2016.

BRASIL, M. d S. Portaria nº 5, de 28 de setembro de 2017. Capitulo V - Dispõe do Controle e da Vigilância da Qualidade da Água para Consumo Humano e seu Padrão de Potabilidade. **Diário Oficial da União**, 2017.

BRENTAN, B. M., LUVIZOTTO Jr, E., MONTALVO, I., IZQUIERDO, J., &PÉREZ-GARCÍA, R. Hybrid regression model for near real-time urban water demand forecasting. **Journal of Computational and Applied Mathematics**, *309*, 532-541, 2017.

BRENTAN, B. M., LUVIZOTTO Jr, E., MONTALVO, I., IZQUIERDO, J., &PÉREZ-GARCÍA, R. Near real time pump optimization and pressure management**.** ***Procedia Engineering***, *186*, 666-675, 2017.

BRENTAN, B., MEIRELLES, G., LUVIZOTTO Jr, E., &IZQUIERDO, J. Joint Operation of Pressure-Reducing Valves and Pumps for Improving the Efficiency of Water Distribution Systems. **Journal of Water Resources Planning and Management,** *144*(9), 04018055. 2018.

BRENTAN, Bruno Melo. Sistemas de abastecimento de água inteligentes: previsão de demanda e controle ótimo em tempo real. 2017.

BRENTAN, Bruno Melo; LUVIZOTTO JR, Edevar; RIBEIRO, Lubienska Cristina LJ. PSO applied to reduce the cost of energy in water supply networks. In: **Applied Mechanics and Materials**. Trans Tech Publications, 2013. p. 703-706.

BROWN, Joe; CAIRNCROSS, Sandy; ENSINK, Jeroen HJ. Water, sanitation, hygiene and enteric infections in children. **Archives of disease in childhood**, v. 98, n. 8, p. 629-634, 2013.

CAO, Li-Juan; TAY, Francis Eng Hock. Support vector machine with adaptive parameters in financial time series forecasting**. IEEE Transactions on neural networks**, v. 14, n. 6, p. 1506-1518, 2003.

CARBONI, D., GLUHAK, A., MCCANN, J. A., &BEACH, T. H. (2016). Contextualising water use in residential settings: A survey of non-intrusive techniques and approaches. **Sensors**, *16*(5), 738.

CARDOSO, S. M., BARROS, D. B., & OLIVEIRA, E. Multiobjective Optimization and Rule Extraction for Optimum Location of Quality Sensors in Water Distribution Network. In **WDSA/CCWI Joint Conference Proceedings**, 2018.

CARDOSO, S. M.; Barros, D. B.; Oliveira, E. C. M.; BRENTAN, B. M.; LUZIVOTTO, E. J.; RIBEIRO, L. C. L. J. Otimização Multiobjetivo para Alocação Ótima de Sensores de Qualidade em Redes de Distribuição de Água. **XXVII Congresso Latinoamericano de Hidráulica**, Buenos Aires, Argentina. 2018a.

CARPITELLA, S., BRENTAN, B., MONTALVO, I., IZQUIERDO, J., & CERTA, A. Multi-criteria analysis applied to multi-objective optimal pump scheduling in water systems. **Water Supply**, 2019.

CASTRO, Pedro; NEVES, Mário. Chlorine decay in water distribution systems case study–lousada network. **Electronic Journal of Environmental, Agricultural and Food Chemistry**, v. 2, n. 2, p. 261-266, 2003.

CEMBRANO, G., WELLS, G., QUEVEDO, J., PÉREZ, R., & ARGELAGUET, R. Optimal control of a water distribution network in a supervisory control system. **Control engineering practice**, *8*(10), 1177-1188, 2010.

CLASEN, Thomas et al. Interventions to improve water quality for preventing diarrhoea. **Cochrane Database of Systematic Reviews**, v. 3, n. 3, p. CD004794, 2006.

CLOETE, Niel Andre; MALEKIAN, Reza; NAIR, Lakshmi. Design of smart sensors for real-time water quality monitoring. **IEEE Access**, v. 4, p. 3975-3990, 2016.

CORTES, Corinna; VAPNIK, Vladimir. Support-vector networks. **Machine learning**, v. 20, n. 3, p. 273-297, 1995.

CRISTO, Cristiana Di; LEOPARDI, Angelo. Pollution source identification of accidental contamination in water distribution networks. **Journal of Water Resources Planning and Management**, v. 134, n. 2, p. 197-202, 2008.

DAS, Subhasish; ROY, Pankaj Kumar; MAZUMDAR, Asis. Analysis of pipe flow and head loss of a modeled network based on EPANET in a water treatment plant at Raipur, West Bengal. **Journal of the Institution of Public Health Engineers, India**, v. 9, n. 3, p. 21-28, 2008.

DAWSEY, Wesley J.; MINSKER, Barbara S.; VANBLARICUM, Vicki L. Reducing online contaminant monitoring uncertainty using a Bayesian belief network. In: **Impacts of Global Climate Change**. 2005. p. 1-12.

DAWSON, C. W.; WILBY, R. L. Hydrological modelling using artificial neural networks. **Progress in physical Geography**, v. 25, n. 1, p. 80-108, 2001.

DEB, Kalyanmoy. Multi-objective optimization. In: **Search methodologies**. Springer, Boston, MA, 2014. p. 403-449.

DENNIS, JR, John E.; MORÉ, Jorge J. Quasi-Newton methods, motivation and theory. **SIAM review**, v. 19, n. 1, p. 46-89, 1977.

DORIGO, Marco et al. (Ed.). Ant Colony Optimization and Swarm Intelligence**. 6th International Conference, ANTS 2008, Brussels, Belgium, September 22-24, 2008, Proceedings**. Springer, 2008.

DREW, Philip J.; MONSON, John RT. Artificial neural networks. **Surgery**, v. 127, n. 1, p. 3-11, 2000.

ELIADES, D. G., KYRIAKOU, M., VRACHIMIS, S., &POLYCARPOU, M. M. EPANET-MATLAB toolkit: An open-source software for interfacing EPANET with MATLAB, 2016.

FISCHER, Thomas; KRAUSS, Christopher. Deep learning with long short-term memory networks for financial market predictions. **European Journal of Operational Research**, v. 270, n. 2, p. 654-669, 2018.

FONSECA, Carlos M.; FLEMING, Peter J. Multiobjective genetic algorithms. In: **Genetic algorithms for control systems engineering, IEE colloquium on**. IET, 1993. p. 6/1-6/5.

FONTANA, Nicola; GIUGNI, Maurizio; PORTOLANO, Davide. Losses reduction and energy production in water-distribution networks. **Journal of Water Resources Planning and Management**, v. 138, n. 3, p. 237-244, 2011.

GARDNER, Matt W.; DORLING, S. R. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. **Atmospheric environment**, v. 32, n. 14-15, p. 2627-2636, 1998.

GHIMIRE, Santosh R.; BARKDOLL, Brian D. A heuristic method for water quality sensor location in a municipal water distribution system: mass-released based approach. In: **Water Distribution Systems Analysis Symposium 2006**. 2008. p. 1-11.

GIACOMELLO, Carlo; KAPELAN, Zoran; NICOLINI, Matteo. Fast hybrid optimization method for effective pump scheduling. **Journal of Water Resources Planning and Management**, v. 139, n. 2, p. 175-183, 2012.

GUNN, Steve R. et al. Support vector machines for classification and regression**. ISIS technical report**, v. 14, n. 1, p. 5-16, 1998.

HAJIBANDEH, Erfan; NAZIF, Sara. Pressure zoning approach for leak detection in water distribution systems based on a multi objective ant colony optimization. **Water resources management**, v. 32, n. 7, p. 2287-2300, 2018.

HALL, John et al. On–Line water quality parameters as indicators of distribution system contamination. **Journal-American Water Works Association**, v. 99, n. 1, p. 66-77, 2007.

HAMED, Maged M.; KHALAFALLAH, Mona G.; HASSANIEN, Ezzat A. Prediction of wastewater treatment plant performance using artificial neural networks. **Environmental Modelling & Software**, v. 19, n. 10, p. 919-928, 2004.

HART, William E.; MURRAY, Regan. Review of sensor placement strategies for contamination warning systems in drinking water distribution systems. **Journal of Water Resources Planning and Management**, v. 136, n. 6, p. 611-619, 2010.

HAYKIN, Simon. **Redes neurais: princípios e prática**. Bookman Editora, 2007.

Herrera, M., Torgo, L., Izquierdo, J., & Pérez-García, R. Predictive models for forecasting hourly urban water demand. **Journal of hydrology**, *387*(1-2), 141-150. 2010.

HU, Xiaohui; EBERHART, Russell. Multiobjective optimization using dynamic neighborhood particle swarm optimization. In: **Evolutionary Computation, 2002. CEC'02. Proceedings of the 2002 Congress on**. Ieee, 2002. p. 1677-1681.

HUTTON, Guy; HALLER, Laurence; BARTRAM, Jamie. Global cost-benefit analysis of water supply and sanitation interventions. **Journal of water and health**, v. 5, n. 4, p. 481-502, 2007.

ILONEN, Jarmo; KAMARAINEN, Joni-Kristian; LAMPINEN, Jouni. Differential evolution training algorithm for feed-forward neural networks**. Neural Processing Letters**, v. 17, n. 1, p. 93-105, 2003.

IZQUIERDO, J., MONTALVO, I., PÉREZ, R., &TAVERA, M. Optimization in water systems: a PSO approach. In **Proceedings of the 2008 Spring simulation multi conference** (pp. 239-246). Society for Computer Simulation International, 2008.

IZQUIERDO, J., LÓPES, P. A., MARTÍNEZ, F. J., & PÉREZ, R. Fault detection in water supply systems using hybrid (theory and data-driven) modelling. **Mathematical and Computer Modelling**, v. 46, n. 3-4, p. 341-350, 2007.

JAIN, Ashu; VARSHNEY, Ashish Kumar; JOSHI, Umesh Chandra. Short-term water demand forecast modelling at IIT Kanpur using artificial neural networks. **Water resources management**, v. 15, n. 5, p. 299-321, 2001.

JANKE, R. MURRAY, R., UBER, J., & TAXON, T. Comparison of physical sampling and real-time monitoring strategies for designing a contamination warning system in a drinking water distribution system. **Journal of water resources planning and management**, v. 132, n. 4, p. 310-313, 2006.

JIANG, Chuanjin; SONG, Fugen. Sunspot Forecasting by Using Chaotic Time-series Analysis and NARX Network. **JCP**, v. 6, n. 7, p. 1424-1429, 2011.

Kennedy, J. and Eberhart, R. C. Particle swarm optimization. Proc. IEEE **Int'l. Conf on Neural Networks**, IV, 1942-1 948. Piscataway, NJ: IEEE Service Center, 1995.

KIRMEYER, Gregory J. **Guidance manual for monitoring distribution system water quality**. American Water Works Association, 2002.

KRAUSE, A., LESKOVEC, J., GUESTRIN, C., VANBRIESEN, J., &FALOUTSOS, C. Efficient sensor placement optimization for securing large water distribution networks. **Journal of Water Resources Planning and Management**, *2008*, 516-526.

KRAUSE, Peter; BOYLE, D. P.; BÄSE, Frank. Comparison of different efficiency criteria for hydrological model assessment. **Advances in geosciences**, v. 5, p. 89-97, 2005.

KREAMER, David K. The past, present, and future of water conflict and international security. **Journal of Contemporary Water Research & Education**, v. 149, n. 1, p. 87-95, 2012.

KULKARNI, Raghavendra V.; VENAYAGAMOORTHY, Ganesh K.; CHENG, Maggie X. Bio-inspired node localization in wireless sensor networks. In: **Systems, Man and Cybernetics, 2009. SMC 2009. IEEE International Conference on**. IEEE, 2009. p. 205-210.

Laird, C. D., Biegler, L. T., van BloemenWaanders, B. G., & Bartlett, R. A. Contamination source determination for water networks. **Journal of Water Resources Planning and Management**, v. 131, n. 2, p. 125-134, 2005.

LEE, Byoung Ho; DEININGER, Rolf A. Optimal locations of monitoring stations in water distribution system. **Journal of Environmental Engineering**, v. 118, n. 1, p. 4-16, 1992.

LEONTARITIS, I. J.; BILLINGS, Stephen A. Input-output parametric models for non-linear systems part I: deterministic non-linear systems. **International journal of control**, v. 41, n. 2, p. 303-328, 1985.

LEVENBERG, Kenneth. A method for the solution of certain non-linear problems in least squares. **Quarterly of applied mathematics**, v. 2, n. 2, p. 164-168, 1944.

LI, Qilin et al. Antimicrobial nanomaterials for water disinfection and microbial control: potential applications and implications. **Water research**, v. 42, n. 18, p. 4591-4602, 2008.

LIU, Li; RANJITHAN, S. Ranji; MAHINTHAKUMAR, G. Contamination source identification in water distribution systems using an adaptive dynamic optimization

procedure. **Journal of Water Resources Planning and Management**, v. 137, n. 2, p. 183-192, 2010.

LIU, M.; WANG, M.; WANG, J., & LI, D. Comparison of random forest, support vector machine and back propagation neural network for electronic tongue data classification: Application to the recognition of orange beverage and Chinese vinegar. **Sensors and Actuators B: Chemical**, v. 177, p. 970-980, 2013.

MAIER, H. R., JAIN, A., DANDY, G. C., &SUDHEER, K. P.Methods used for the development of neural networks for the prediction of water resource variables in river systems: current status and future directions. **Environmental modelling & software**, v. 25, n. 8, p. 891-909, 2010.

MAIER, Holger R.; DANDY, Graeme C. The use of artificial neural networks for the prediction of water quality parameters. **Waterresourcesresearch**, v. 32, n. 4, p. 1013-1022, 1996.

MEIRELLES, Gustavo et al. Trunk network rehabilitation for resilience improvement and energy recovery in water distribution networks. **Water**, v. 10, n. 6, p. 693, 2018.

MARQUARDT, Donald W. An algorithm for least-squares estimation of nonlinear parameters. **Journal of the society for Industrial and Applied Mathematics**, v. 11, n. 2, p. 431-441, 1963.

MENEZES JR, José Maria P.; BARRETO, Guilherme A. Long-term time series prediction with the NARX network: an empirical evaluation. **Neurocomputing**, v. 71, n. 16-18, p. 3335-3343, 2008.

MENEZES JR, Jose MP; BARRETO, Guilherme A. A new look at nonlinear time series prediction with NARX recurrent neural network. In: **2006 Ninth Brazilian Symposium on Neural Networks (SBRN'06)**. IEEE, 2006. p. 160-165.

MENG, X. B., GAO, X. Z., Lu, L., LIU, Y., &ZHANG, H. (2016). A new bio-inspired optimisation algorithm: Bird Swarm Algorithm. **Journal of Experimental & Theoretical Artificial Intelligence,** *2016*, 673-687.

MITCHELL, Melanie. **An introduction to genetic algorithms**. MIT press, 1998.

MONTALVO, I., IZQUIERDO, J., PÉREZ, R., &TUNG, M. M. Particle swarm optimization applied to the design of water supply systems. **Computers & Mathematics with Applications**, *2008*, 769-776.

MONTELEONE, M. C.; YEUNG, Hoi; SMITH, Richard. A review of Ancient Roman water supply exploring techniques of pressure reduction. **Water Science and Technology: Water Supply**, v. 7, n. 1, p. 113-120, 2007.

MORÉ, Jorge J. The Levenberg-Marquardt algorithm: implementation and theory. In: **Numerical analysis**. Springer, Berlin, Heidelberg, 1978. p. 105-116.

MOZAFFARI, Ahmad; FATHI, Alireza; BEHZADIPOUR, Saeed. The great salmon run: a novel bio-inspired algorithm for artificial system design and optimisation. **International Journal of Bio-Inspired Computation**, v. 4, n. 5, p. 286-301, 2012.

MUKHERJEE, Rajib; DIWEKAR, Urmila M.; VASEASHTA, Ashok. Optimal sensor placement with mitigation strategy for water network systems under uncertainty. **Computers & Chemical Engineering**, v. 103, p. 91-102, 2017.

MØLLER, Martin Fodslette. A scaled conjugate gradient algorithm for fast supervised learning. **Neural networks**, v. 6, n. 4, p. 525-533, 1993.

NAZIF, Sara et al. Pressure management model for urban water distribution networks. **Water resources management**, v. 24, n. 3, p. 437-458, 2010.

NEUMANN, Peter; NEUMANN, Peter R. **Old and new terrorism**. Polity, 2009.

NICOLINI, Matteo; ZOVATTO, Luigino. Optimal location and control of pressure reducing valves in water networks. **Journal of water resources planning and management**, v. 135, n. 3, p. 178-187, 2009.

NOBLE, William S. What is a support vector machine?.**Nature biotechnology**, v. 24, n. 12, p. 1565, 2006.

OHAR, Ziv; LAHAV, Ori; OSTFELD, Avi. Optimal sensor placement for detecting organophosphate intrusions into water distribution systems. **Water research**, v. 73, p. 193-203, 2015.

OLIVEIRA, Eva Carolline M. et al. Detection of chemical intrusion compounds in water distribution networks by quality sensors data mining. In: **WDSA/CCWI Joint Conference Proceedings**. 2018.

Ostfeld, A., Uber, J. G., Salomons, E., Berry, J. W., Hart, W. E., Phillips, C. A., ... & di Pierro, F. The battle of the water sensor networks (BWSN): A design challenge for engineers and algorithms. **Journal of Water Resources Planning and Management**, *134*(6), 556-568, 2008.

OSTFELD, Avi; SALOMONS, Elad. Sensor network design proposal for the battle of the water sensor networks (BWSN). In: **Water Distribution Systems Analysis Symposium 2006**. 2008. p. 1-16.

PAES, A. C., da SILVA, M. B. P., & GOMES, E. N. Estimativa da Irradiação Solar Global pelos Modelos de Hargreaves-Samani e Aprendizado de Máquina SVM e ANN em Botucatu/SP/Brasil. In: **VII Congresso Brasileiro de Energia Solar-CBENS 2018**. 2018.

PALLETI, Venkata Reddy et al. Sensor network design for contaminant detection and identification in water distribution networks. **Computers & Chemical Engineering**, v. 87, p. 246-256, 2016.

PÉREZ, Carlos J. et al. A Multi-Objective Artificial Bee Colony-based optimization approach to design water quality monitoring networks in river basins. **Journal of Cleaner Production**, v. 166, p. 579-589, 2017.

PING, Tristan Sim Tong. Terrorism—a new perspective in the water management landscape. **International Journal of Water Resources Development**, v. 26, n. 1, p. 51-63, 2010.

PREIS, Ami; OSTFELD, Avi. Multiobjective contaminant sensor network design for water distribution systems. **Journal of Water Resources Planning and Management**, v. 134, n. 4, p. 366-377, 2008.

RAO, Zhengfu; BERMEJO, Fernando Alvarruiz. Use of an artificial neural network to capture the domain knowledge of a conventional hydraulic simulation model. In: **Journal of Hydroinformatics**. IWA Publishing, 2007. p. 15-24.

RASEKH, A., HASSANZADEH, A., MULCHANDANI, S., MODI, S., & BANKS, M. K. Smart water networks and cyber security. **Water Resources Planning and Management**. 2016
REY HORN, Je; NAFPLIOTIS, Nicholas; GOLDBERG, David E. Multiobjective optimization using the niched pareto genetic algorithm. **IlliGAL report**, v. 93005, 1993.

RIBEIRO, L. C. L. J. "Modelo Híbrido Multiobjetivo para Obtenção de Roteiros Operacionais de Bombas de Rotação Variável em Instalações Hidráulicas". Campinas: University of Campinas (Unicamp), Brazil, 2007.

ROGERS, Katrina S. Pre-empting violent conflict: learning from environmental cooperation. In: **Conflict and the Environment**. Springer, Dordrecht, 1997. p. 503-518.

ROSSMAN, L.A. EPANET Users Manual, EPA-600/R-94/057, Risk Reduction Engineering Laboratory, U.S. **Environmental Protection Agency**, Cincinnati, OH. 1994.

ROSSMAN, Lewis A. The EPANET programmer's toolkit for analysis of water distribution systems. In: **WRPMD'99: Preparing for the 21st Century**. 1999. p. 1-10.

RUMELHART, David E.; HINTON, Geoffrey E.; WILLIAMS, Ronald J. Learning internal representations by error propagation. California Univ San Diego La Jolla Inst for **Cognitive Science**, 1985.

SALOMONS, Elad et al. Optimizing the operation of the Haifa-A water-distribution network. **Journal of Hydroinformatics**, v. 9, n. 1, p. 51-64, 2007.

SANZ, Gerard; PÉREZ, Ramon. Benchmark de control y supervisión de redes de distribución de agua. **Actas de las XXXIV Jornadas de Automática**, p. 1-6, 2013.

SCHUSTER, Mike; PALIWAL, Kuldip K. Bidirectional recurrent neural networks. **IEEE transactions on Signal Processing**, v. 45, n. 11, p. 2673-2681, 1997.

SHEPHERD, Adrian J. Second-order methods for neural networks: Fast and reliable training methods for multi-layer perceptrons. **Springer Science & Business Media**, 2012.

SHANG, F., UBER, J. G., ROSSMAN, L. A., JANKE, R., &MURRAY, R. EPANET multi-species extension user's manual. *Risk Reduction Engineering Laboratory,* **US Environmental Protection Agency***, Cincinnati, Ohio*. 2008.

SHI, Yuhui; EBERHART, Russel C. Particle swarm optimization: developments, applications and resources. In: **evolutionary computation, 2001. Proceedings of the 2001 Congress on**. IEEE, 2001. p. 81-86.

SHMILOVICI, Armin. Support vector machines. In: **Data mining and knowledge discovery handbook. Springer**, Boston, MA, 2009. p. 231-247

STAFF, Joint Chiefs Of. **Department of Defense dictionary of military and associated terms**. MILITARY BOOKSHOP, 2013.

STONE, R. J. Improved statistical procedure for the evaluation of solar radiation estimation models. **Solar energy**, v. 51, n. 4, p. 289-291, 1993.

STOREY, Michael V.; VAN DER GAAG, Bram; BURNS, Brendan P. Advances in on-line drinking water quality monitoring and early warning systems. **Water Research**, v. 45, n. 2, p. 741-747, 2011.

SUSE, Roberto; JUNIOR, Edevar Luvizotto; BRENTAN, Bruno Melo. Modelo Matemático para Identificar Locais de Estações de Monitoramento em Redes de Distribuição de Água.**XII Simposio Iberoamericano sobre planificación de sistemas de abastecimiento y drenaje.**2013.

SUSE, Roberto; YOSHIKAWA, Andréa Manami; JUNIOR, EdevarLuvizotto. Simulação da qualidade de água em redes de distribuição empregando o EPANET-MSX. **Revista DAE**, n. 197, p. 32-41, 2014.

Witten, I. H., Frank, E., Hall, M. A., &Pal, C. J. **Data Mining: Practical machine learning tools and techniques**. Morgan Kaufmann, 2016

WRIGHT, Jim; GUNDRY, Stephen; CONROY, Ronan. Household drinking water in developing countries: a systematic review of microbiological contamination between source and point-of-use. **Tropical medicine & international health**, v. 9, n. 1, p. 106-117, 2004.

WU, Yipeng; LIU, Shuming. A review of data-driven approaches for burst detection in water distribution systems. **Urban Water Journal**, v. 14, n. 9, p. 972-983, 2017.

Yaghlane, Asma Ben, and M. Naceur Azaiez. "Systems under attack-survivability rather than reliability: Concept, results, and applications." **European Journal of Operational Research,** 258, 2017.

Zhang G.P. Neural Networks for Data Mining. In: Maimon O., Rokach L. (eds) **Data Mining and Knowledge Discovery Handbook**. Springer, Boston, MA, 2009.

Zhe, X.; Jie, Y.; Huaqiang, C.; Yaguang, K. &Bishi, H. Water Distribution NetworkModeling Based on NARX. **IFAC**. Vol. 48. 2015.

ZULKIFLI, Syahidah Nurani; RAHIM, Herlina Abdul; LAU, Woei Jye. Detection of Contaminants in Water Supply: A Review on State-of-the-Art Monitoring Technologies and Their Applications. **Sensors and Actuators B: Chemical**, 2017.

Apêndices

A.1. Simulação da Qualidade da Água

Essa etapa divide-se entre as aplicações das metodologias de Cardoso *et al*. (2018), Oliveira *et al*. (2018) e EPANET-MATLAB Toolkit no qual se obtém o banco de dados para a aplicação nos métodos de aprendizado de máquinas.

A.1.1. Alocação de Sensores

Para a alocação de sensores e identificação da contaminação, é necessária uma simulação em que se considere todos os nós como possíveis pontos de ataques e possíveis locais de sensores. Para isso, aplicou-se, em cada nó, a concentração de 12,4 mg/L do pesticida Paration, concentração também utilizada na pesquisa de Oliveira *et al*. (2018), e houve a análise a cada 600 segundos das interações do pesticida em todos os nós, incluídos os nós utilizados como sensores.

Como tempo de simulação adotou-se 48 horas, na qual há a injeção de cloro, a uma concentração de 1,5 ml/L, em todos os reservatórios durante todo tempo simulado e, após 24 horas de simulação, ocorre a injeção do pesticida durante 9 horas. Esse processo se repete, nesse caso, 15 vezes, visto ser a quantidade de nós presente na RDA estudada, esse processo segue a metodologia proposta por Cardoso *et al*. (2018).

Após essa simulação, aborda-se o algoritmo NSGA-II, em que segue as funções objetivo Z1 e Z4 da BWSN, para a alocação de dois sensores na RDA, no qual gera os resultados otimizados dessa alocação e apresenta uma Fronteira de Pareto (Figura A.1), com os possíveis conjuntos de soluções.

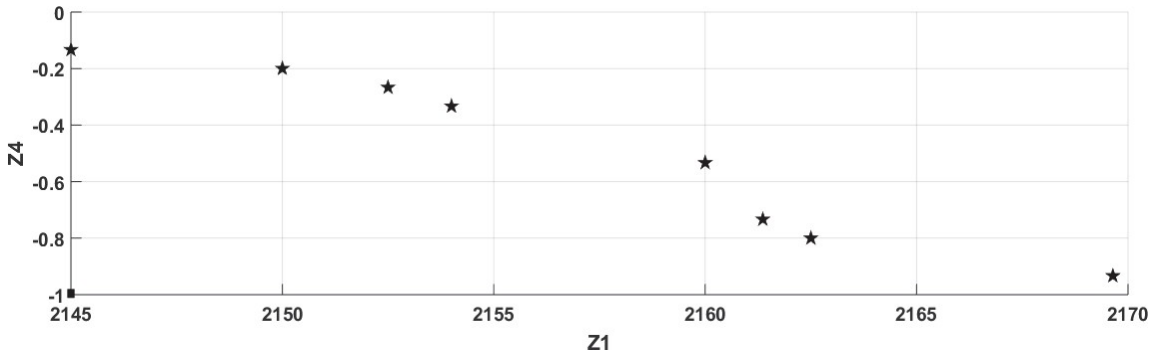

Figura A.1 – Fronteira de Pareto - RZM.

Para a escolha do melhor conjunto de soluções, utilizou-se como extrator de regra a menor distância euclidiana, ou seja, a menor distância entre os conjuntos a um ponto ideal, neste caso representado na Figura A.1, como um ponto e os conjuntos representados como estrelas.

Como resultado, o conjunto que contém os nós 12 e 15, é selecionado. Visto que o nó 15 é apontado como melhor ponto de monitoramento, uma vez que recebe contribuição de água de todos os pontos da RDA. Já, a alocação de sensor no nó 12, permite a redução no tempo de detecção de possíveis contaminações químicas.

A alocação de sensores para a rede JYN, Figura A.2, seguiu a mesma metodologia apresentada, sendo a quantidade de sensores estipulada igual a 6. Neste processo ocorreram 302 cenários de contaminação, sendo incluídos os dos reservatórios como possíveis pontos para alocação de sensores.

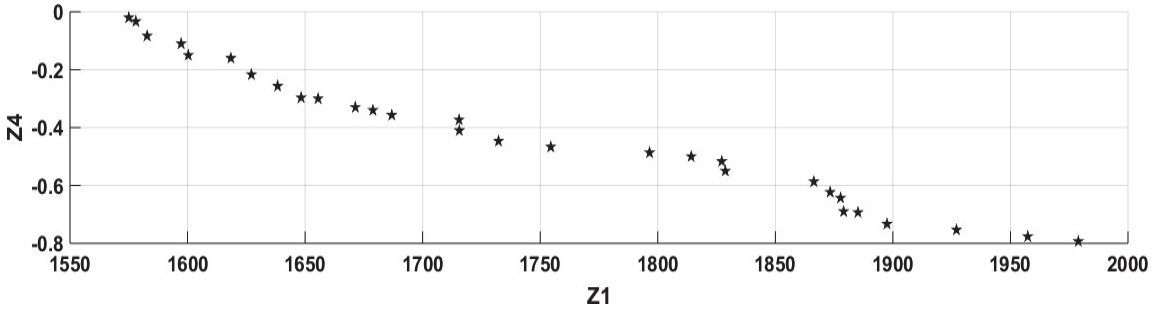

Figura A.2 – Fronteira de Pareto - JYN.

Os nós selecionados para a alocação de sensores foram 53, 30, N7, N272, N266 e N44. Seguindo a pesquisa de He *et al.* (2018), que indicaram a alocação de 6 sensores, dos quais dois também foram selecionados nesta pesquisa.

A.2. Avaliação de erros das RNs para definição da arquitetura ótima

A variação no número de neurônios por camadas ocultas permitiu a verificação na quantidade de erros apresentados por cada arquitetura. Analisou-se a variação dos erros ocasionados pela mudança na quantidade de neurônios das duas primeiras camadas. Essas variações são apresentadas nas Figuras A.3 e A.4 para as aplicações de FIT e FEED, respectivamente.

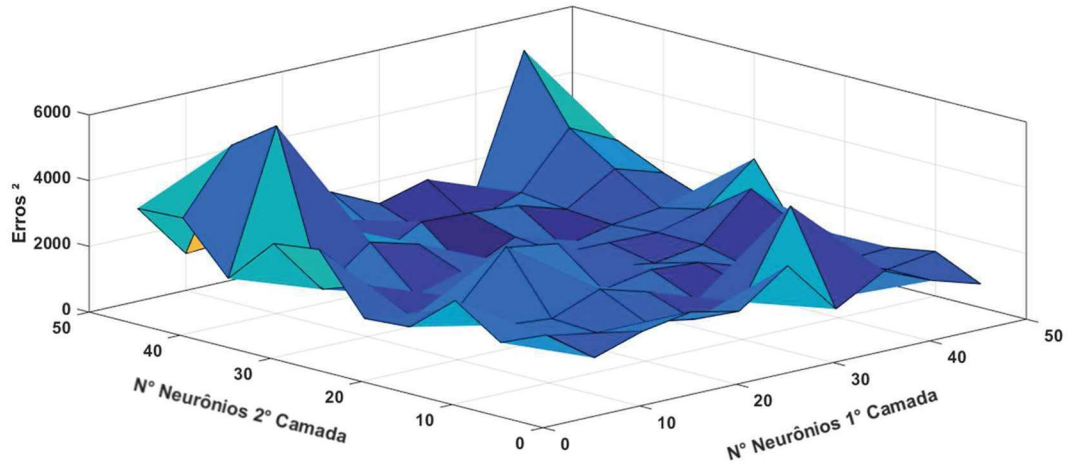

Figura A.3–Erros pelos números de neurônios – FIT

Observa-se, na Figura A.3, que a aplicação FIT apresenta grandes erros quando utilizados entre 0 e 20 neurônios na primeira camada oculta entre 40 e 50 na segunda camada. Isso se repete, porém com maior intensidade, quando utilizados altos valores de neurônios em ambas as camadas. A configuração escolhida, como dito anteriormente, possui 10 na primeira camada e 40 na segunda.

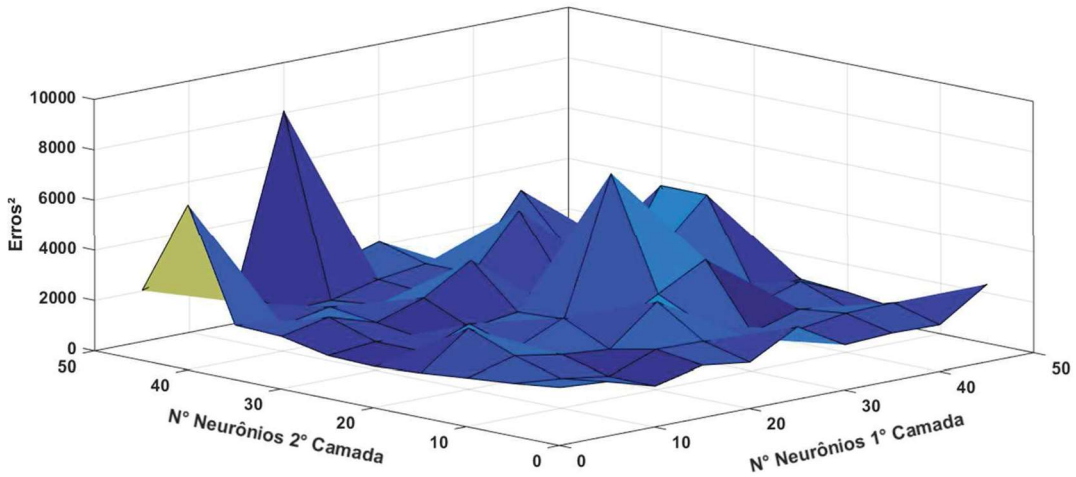

Figura A.4 - Erros pelos números de neurônios – FEED

Na aplicação de FEED, os maiores erros são encontrados quando utilizados entre 10 - 20 neurônios na primeira camada e aproximadamente 50 neurônios na segunda camada. Pode-

se observar, ainda, altos valores de erros quando utilizado valores entre 20 e 30 de neurônios em ambas as camadas.