UFBA
UNIVERSIDADE FEDERAL DA BAHIA
ESCOLA POLITÉCNICA
PROGRAMA DE PÓS GRADUAÇÃO EM
ENGENHARIA INDUSTRIAL - PEI

MESTRADO EM ENGENHARIA INDUSTRIAL

BRENNER BIASI SOUZA SILVA

MINERAÇÃO DE DADOS PARA PREDIÇÃO DE FALHA EM SISTEMA DE COLETA DE EFLUENTES

SALVADOR
2019

UNIVERSIDADE FEDERAL DA BAHIA
ESCOLA POLITÉCNICA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA INDUSTRIAL

BRENNER BIASI SOUZA SILVA

MINERAÇÃO DE DADOS PARA PREDIÇÃO DE FALHA EM SISTEMA DE COLETA DE EFLUENTES

Salvador
2019

BRENNER BIASI SOUZA SILVA

MINERAÇÃO DE DADOS PARA PREDIÇÃO DE FALHA EM SISTEMA DE COLETA DE EFLUENTES

Dissertação de Mestrado Acadêmico apresentada ao Programa de Pós-graduação em Engenharia Industrial, da Universidade Federal da Bahia, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia Industrial.

Orientadores: Prof. DSc. Karla Patrícia Santos Oliveira Rodriguez Esquerre
MSc. Robson Wilson Silva Pessoa

Salvador
2019

Ficha catalográfica elaborada pelo Sistema Universitário de Bibliotecas (SIBI/UFBA),
com os dados fornecidos pelo(a) autor(a).

Silva, Brenner Biasi Souza
Mineração de Dados para Predição de Falha em Sistema de Coleta de Efluentes / Brenner Biasi Souza Silva. -- Salvador, 2019.
96 f. : il

Orientadora: Karla Patricia Santos Oliveira-Esquerre.
Coorientador: Robson Wilson Silva Pessoa.
Dissertação (Mestrado - Programa de Pós-Graduação em Engenharia Industrial) -- Universidade Federal da Bahia, Escola Politécnica, 2019.

1. Transbordamento de efluentes industriais. 2. Aprendizado não supervisionado. 3. Aprendizado supervisionado. I. Oliveira-Esquerre, Karla Patricia Santos. II. Pessoa, Robson Wilson Silva. III. Título.

MINERAÇÃO DE DADOS PARA PREDIÇÃO DE FALHA EM SISTEMA DE COLETA DE EFLUENTES

BRENNER BIASI SOUZA SILVA

Dissertação submetida ao corpo docente do programa de pós-graduação em Engenharia Industrial da Universidade Federal da Bahia como parte dos requisitos necessários para a obtenção do grau de mestre em Engenharia Industrial.

Examinada por:

Profa. Dra. Karla Patricia Santos Oliveira Rodríguez Esquerre _____
Doutora em Engenharia Química pela Universidade Estadual de Campinas, Brasil, 2003.

Prof. Dr. Luciano Matos Queiroz _____
Doutor em Engenharia Civil pela Universidade de São Paulo, Brasil, 2009.

Prof. Dr. Leizer Schnitman _____
Doutor em Engenharia Eletrônica e Computação, pelo Instituto Tecnológico de Aeronáutica, Brasil, 2001.

Salvador, BA - BRASIL
Junho/2019

AGRADECIMENTOS

Agradeço aos meus pais, José Ubaldo Gonzaga da Silva e Mércia de Souza Silva, por todo amor, suporte, ensinamentos e incentivo em busca do crescimento técnico-científico.

A Geise Araújo por todo companheirismo, paciência, afeto e sabedoria que foram necessários em cada momento para vencer este desafio.

A minha orientadora Karla Oliveira Esquerre pela confiança, incentivo, inspiração, compartilhamento de conhecimentos e suporte acadêmico, principalmente em momentos incertos.

A Robson Pessoa pela disponibilidade, presteza, compartilhamento de ideias, intensidade e generosidade, apresentando e encaminhando informações técnicas e científicas.

A Carla Pereira por todo suporte técnico e colaboração necessária para pleno desenvolvimento da pesquisa.

A Gabriela Botelho, professora e amiga de laboratório, pela confiança e contribuição para os primeiros passos dessas caminhada.

Aos membros/amigos do grupo GAMMA, pelo compartilhamento de conhecimentos, solidariedade, apoio e pelos excepcionais momentos proporcionados durante esta jornada.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

SILVA, Brenner Biasi Souza. Mineração de dados para predição de falhas em sistemas de coleta de efluentes. 2019. Orientadores: Karla Patrícia Santos Oliveira Rodriguez Esquerre e Robson Wilson Silva Pessoa. 92 f. il. Dissertação (Mestrado em Engenharia Industrial) – Escola Politécnica, Universidade Federal da Bahia, Salvador, 2019.

RESUMO

A ocorrência de transbordamentos nos sistemas de retenção de efluentes industriais é um aspecto ambiental e operacional importante na maioria das indústrias. O objetivo deste trabalho foi analisar o comportamento do sistema de coleta e contenção de efluentes industriais em uma refinaria de petróleo, e propor um modelo preditivo para evento de falha, transbordamento, de uma bacia de contenção. A análise inicial foi realizada utilizando técnica de agrupamento para objetos considerando Índice de Similaridade, sendo também realizado abordagem de séries temporais de precipitação pluviométrica e percentual do nível de tanques de contenção do sistema a partir da perspectiva de similaridade, da detecção de pontos de mudança e análise de tendências. Modelos preditivos foram construídos utilizando *k-nearest neighbors* (KNN) e *Random Forest* para predição de classificação, com o objetivo de indicar se a bacia de contenção transbordará numa projeção para o horizonte de 24 horas. O conjunto de metodologias de aprendizado de máquina não supervisionadas usadas aqui permite obter informações sobre eventos hidrológicos e de processo em cenários com baixa disponibilidade de dados sem a necessidade de aumentar a informação. Identificou-se que, na ausência de precipitação ou ocorrência de baixos volumes diários de precipitação, o sistema falhou, e a porcentagem de transbordamentos é maior do que o valor natural esperado. Além disso, não houveram transbordamentos em períodos chuvosos em casos de operação considerada satisfatória do sistema. Cenários e variações de técnicas de amostragem para o treinamento dos modelos de classificação foram utilizados. Os melhores resultados dos modelos preditivos construídos foram obtidos a partir do algoritmo *Random Forest* com emprego da técnica de reamostragem *oversampling*, *undersampling* e ROSE.

Palavras-Chave: Transbordamento de efluentes industriais. Aprendizado não-supervisionado. Aprendizado supervisionado.

SILVA, Brenner Biasi Souza. Data mining for failure prediction in wastewater collection systems. 2019. Thesis advisor: Karla Patrícia Santos Oliveira Rodriguez Esquerre e Robson Wilson Silva Pessoa. 92 f. il Dissertation (Master in Industrial Engineering) – Escola Politécnica, Universidade Federal da Bahia, Salvador, 2019.

ABSTRACT

The occurrence of overflows in industrial effluent retention systems is an important environmental and operational aspect in most industries. The aim of this work was to analyze the behavior of the industrial effluent collection and retention system in an oil refinery, and to propose a predictive model for the event of failure, overflow, of a retention tank. The initial analysis was carried out using a clustering technique for objects considering the Similarity Index, and also an approach of time series of rainfall and percentage of the level of containment tanks of the system from the perspective of similarity, the detection of change points and analysis of trends. Predictive models were constructed using k-nearest neighbors (KNN) and Random Forest for classification prediction, with the objective of indicating whether the retention tank will overflow in a 24-hour horizon projection. The set of unsupervised machine learning methodologies used here allows to obtain information on hydrological and process events in scenarios with low data availability without the need to increase the information. It was identified that, in the absence of precipitation or occurrence of low daily precipitation volumes, the system failed, and the percentage of overflows is higher than the expected natural value. In addition, there were no overflows during rainy periods in cases where the system's operation was considered satisfactory. Scenarios and variations in sampling techniques for training the classification models were used. The best results of the constructed predictive models were obtained from the Random Forest algorithm using the oversampling, undersampling and ROSE re-sampling techniques.

Keywords: Overflow of industrial effluents. Unsupervised learning. Supervised learning.

SUMÁRIO

1 INTRODUÇÃO 8
1.1 CONSIDERAÇÕES INICIAIS 8
1.2 OBJETIVOS DA PESQUISA 10
1.3 ESTRUTURA DA DISSERTAÇÃO 11
2 REVISÃO BIBLIOGRÁFICA 13
2.1 ASPECTOS HIDROLÓGICOS 13
2.1.1 Tratamento de Dados Hidroclimatológicos 15
2.2 SISTEMAS DE DRENAGEM 17
2.3 APRENDIZADO DE MÁQUINA 21
2.3.1 Aprendizado Não Supervisionado 23
2.3.1.1 Agrupamento de objetos 23
2.3.1.1.1 Métodos de validação de agrupamentos 25
2.3.1.2 Change-point 28
2.3.2 Aprendizado Supervisionado 30
2.3.2.1 K-nearest neighbors - KNN 31
2.3.2.2 Métodos baseados em árvores 32
2.3.2.3 Métodos de amostragem 37
2.3.2.4 Validação de Modelos 38
2.3.2.5 Avaliação de performance de modelos 39
3 METODOLOGIA 42
4 OBJETO DE ESTUDO 42
4.1 TRATAMENTO DE DADOS 45
4.2 ANÁLISE DE COMPORTAMENTO E PADRÕES DE TRANSBORDAMENTO 45
4.2.1 Índice de similaridade para agrupamentos 45
4.2.2 *Change-point* 48
4.3 MODELAGEM PREDITIVA PARA TRANSBORDAMENTO 49
4.3.1 Avaliação dos modelos de predição 52
5 RESULTADOS 53
5.1 ANÁLISE EXPLORATÓRIA 53
5.2 ANÁLISE DE COMPORTAMENTO DE OPERAÇÃO DAS BACIAS 58
5.3 MODELAGEM PREDITIVA DE TRANSBORDAMENTO 67
6 CONCLUSÃO 72

7 SUGESTÃO PARA TRABALHOS FUTUROS 74
REFERÊNCIAS 75
APÊNDICES - A 82

1 INTRODUÇÃO

Este capítulo diz respeito à motivação, contexto e objetivos da presente dissertação, bem como a introdução do trabalho de pesquisa desenvolvido.

No primeiro momento são apresentadas problemáticas ambientais acerca do evento de transbordamento de efluentes industriais, com foco em refinarias, onde as relações de causa e efeito são introduzidas, e o evento de transbordamento é conceituado como falha. Diferentes abordagens de análise e predição de falhas são citadas, entretanto as técnicas de mineração de dados são destacadas e exemplificadas ao ponto da utilidade e do ganho científico.

Em seguida, as principais questões relacionadas às contribuições deste trabalho são destacadas através dos objetivos alcançados. Por fim, a estrutura dos próximos capítulos da dissertação é apresentada.

1.1 CONSIDERAÇÕES INICIAIS

A conservação dos recursos naturais é uma das soluções para a sustentabilidade. Estudos recentes sinalizam grande preocupação quanto a real ou potencial contaminação e poluição oriunda de atividades industriais ao ambiente, principalmente ocasionadas por refinarias de petróleo (DIYA'UDDEEN; DAUD; ABDUL, 2011; MARCUS; EKPETE, 2014).

Tais atividades industriais envolvem o processamento de compostos químicos perigosos, que a priori representam risco potencial ao ambiente pela possibilidade de serem inadequadamente lançados no ambiente (OLIVEIRA-ESQUERRE et al., 2011; OSIN; YU; LIN, 2017).

O manejo adequado dos efluentes minimiza o perigo e risco inerente ao meio. Neste contexto, os sistemas de contenção de efluentes são de grande importância. Estes sistemas funcionam como barreiras de segurança operacional dos sistemas de drenagem na qual o efluente é armazenado em bacias por determinado período, sendo as bacias operadas de acordo a vazão de entrada, a capacidade hidrodinâmica do sistema e característica do efluente.

O funcionamento dos sistemas de contenção de efluentes industriais é similar a outros sistemas de drenagem, como o de drenagem urbana, principalmente quando o

sistema de drenagem é combinado com o de esgotamento sanitário, como pode ser notado em Martino et al. (2011), pois ambos os sistemas podem transbordar em situação com efluentes em excesso na estrutura para contenção de efluentes.

Dentre as possíveis análises para o sistema, a formulação de relações de causa e efeito relacionada aos comportamentos regulares e anômalos das bacias de contenção podem ser explorados à luz dos processos de chuva-vazão, tendo em vista os possíveis efeitos colaterais acerca do ciclo hidrológico e características da bacia hidrográfica (MAILHOT; TALBOT; LAVALLÉE, 2015; SCHROEDER et al., 2011).

Os eventos de falhas em sistemas de drenagem urbana podem ser conceituados como ocorrências de sobrevazão, inundação ou transbordamentos (THORNDAHL; SCHAARUP-JENSEN; JENSEN, 2008). Soluções como abordagens mais clássicas de diagnóstico de falhas (VENKATASUBRAMANIAN; RENGASWAMY; YIN, 2003), bem como aplicação de técnica de controle (OCAMPO-MARTINEZ, 2010) e simulação hidrológica (ARTINA et al., 2007) também podem ser um caminho para analisar e solucionar esta problemática relacionada ao transbordamento de efluentes.

No contexto industrial, a gestão dos sistemas de drenagem é aperfeiçoada pelo monitoramento, a análise de comportamento e o reconhecimento de padrões dos seus efluentes. Estas ações são muito importantes para melhorar a eficiência, a produtividade e a predição dos processos, inclusive podendo ser agentes facilitadores para a tomada de decisão (HU; CHEN; SHAH, 2018; JYOTI; SINGH, 2011; LEITÃO; AFFONSO GUEDES, 2016).

O avanço científico tecnológico relacionado a sistemas de informações aumentam a demanda por sistemas que possibilitem a extração de informações de maneira rápida e eficiente. Entretanto, os desafios para esta implementação são o pensamento crítico, a avaliação de recursos científicos e a possibilidade de aproveitamento da infraestrutura disponível (GE et al., 2017; SAUCEDO-MARTÍNEZ et al., 2018).

Neste cenário modernista, a Indústria 4.0, desenvolvida para grandes massas de dados e com dispositivos e sistemas inteligentes, tem potencial para reduzir desperdícios e proporcionar uma indústria mais sustentável (KAMBLE; GUNASEKARAN; GAWANKAR, 2018).

Em sistemas vulneráveis a eventos de precipitação pluviométrica (sistemas a céu aberto) que possam ser considerados como importantes ou gargalos, a análise de

comportamento e caracterização do evento de causa e efeitos são essenciais para o reconhecimento de padrões operacionais, permitindo maior conhecimento acerca desta relação, como estudado por Löwe; Madsen e McSharry (2016), Scholz (2008) e Yu et al. (2013, 2018).

Resultados de estudos das relações de causa e efeito oferecem como subsídio uma base de informações preliminares para orientar os gestores na sua avaliação do planejamento e dos riscos em vários cenários operacionais (LIU et al., 2016).

A construção de modelos preditivos, o desenvolvimento de conhecimento baseado em modelagem empírica, utilizando algoritmos computacionais e metodologia como mineração de dados e aprendizado de máquina, são importantes na descoberta de conhecimento e na tomada de decisões nas diversas possíveis áreas de aplicação (GE et al., 2017).

Técnicas de aprendizado de máquina tem propiciado mais informações de modo a facilitar a tomada de decisão. Tais métodos apresentam grande potencial para análise de dados sendo aplicados com sucesso em várias áreas da engenharia e podem ser cada vez mais requisitados, tendo em vista os avanços no desenvolvimento de sensores e sistemas de transmissão da informação (SYAFRUDIN et al., 2018).

1.2 OBJETIVOS DA PESQUISA

Trabalhos desenvolvidos utilizando técnicas de mineração de dados para melhor compreensão sobre eventos e desenvolvimento de soluções em sistemas de drenagem urbana e processos industriais tem crescido nos últimos anos.

Temas envolvendo sistemas de drenagem industriais utilizando mineração de dados ainda foram pouco explorados na literatura. Entre estes temas destacam-se: análise de similaridade entre eventos de transbordamentos e características da precipitação pluviométrica através de métodos não-supervisionados em situações com dados de eventos hidrológicos com baixa qualidade; desenvolvimento de atributos em etapa de pré-processamento de dados a partir de considerações acerca do fenômeno e do sistema de engenharia em análise; predição de falha, transbordamento, em tanque de contenção.

A partir da revisão de literatura e da avaliação de semelhanças entre fenômenos urbanos e industriais, é perceptível lacunas científicas acerca da modelagem empírica

sobre sistemas de drenagem de efluentes industriais, principalmente voltado para predição de falhas, isto é, transbordamentos.

Nesta dissertação pretende-se construir modelo preditivo de classificação para a ocorrência de falhas, fenômeno de transbordamento, de um sistema de coleta e contenção de drenagem de efluentes industrias em uma refinaria de petróleo. Para tal, com o propósito de realização de pesquisa aplicada, objetivos específicos foram consolidados de tal forma que:

- Atributos foram desenvolvidos no pré-processamento de dados a partir de variáveis medidas em campo;
- Análises baseadas em índice de similaridade considerando agrupamentos de dados de chuva e operação das bacias de contenção nortearam a necessidade de novas análises a partir de outro método;
- O perfil de comportamento das séries de chuva e nível das bacias de contenção do sistema de drenagem foi analisado de forma a complementar a análise de índice de similaridade; e
- Diferentes modelos preditivos baseados em classificação quanto à ocorrência ou não de transbordamento na principal bacia de contenção do sistema de drenagem foram construídos e analisados quanto a performance.

1.3 ESTRUTURA DA DISSERTAÇÃO

A dissertação está dividida em oito capítulos e um apêndice, estruturados da seguinte forma:

O capítulo 1 (presente capıtulo) trata da introdução e dos objetivos deste trabalho.

No capítulo 2 é apresentado a revisão do estado da arte, onde são apresentados os aspectos teóricos e práticos relacionados a situações de causa e efeito, envolvendo a hidrologia e fenômenos de transbordamentos, para sistemas de drenagem urbana, que são analisados e exemplificados para posterior analogia a sistemas de drenagem industrial. Técnicas de mineração de dados também são exemplificadas tal que potencialidades e desvantagens sejam apresentadas.

O capıtulo 3 apresenta a metodologia utilizada para o presente trabalho .

No capítulo 4 o objeto de estudo é apresentado detalhadamente, elucidando atributos, o tipo e qualidade dos dados, aspectos operacionais e características do sistema de drenagem de efluentes industriais. Também são explanados os métodos aplicados para mineração de dados não-supervisionada e supervisionada, e as variáveis utilizadas.

As conclusões, são apresentadas no capítulo 5. Indicações para futuros trabalhos a serem desenvolvidos nesta linha de pesquisa são apresentados no capítulo 6.

No apêndice são apresentados gráficos e tabelas que foram utilizados como suporte para discussão dos resultados e conclusão.

2 REVISÃO BIBLIOGRÁFICA

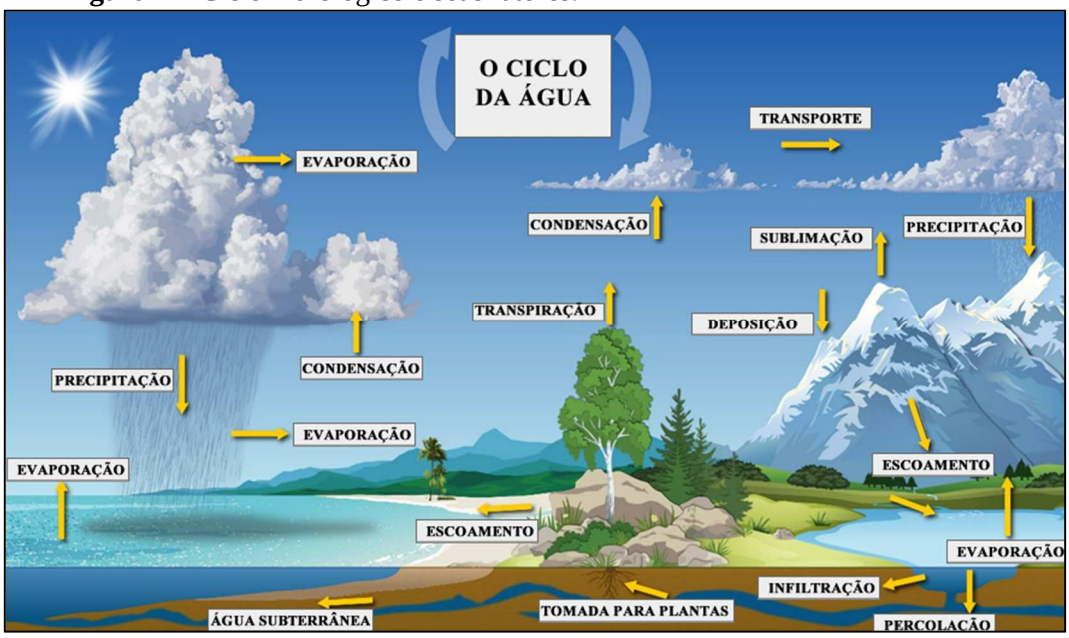
2.1 ASPECTOS HIDROLÓGICOS

Estudos hidrológicos buscam avaliar a ocorrência, circulação, distribuição espacial da água na Terra, bem como suas propriedades físicas e químicas e sua relação com o ambiente.

O regime hidrológico de uma região é determinado por suas características físicas, topográficas e clima local, e os fatores climáticos de suma importância são a precipitação, de acordo com a distribuição espacial e modos de ocorrência, e a evaporação.

Os fatores climáticos em questão são alguns dos componentes do ciclo hidrológico, que é um fenômeno global de circulação fechada da água entre a superfície terrestre e a atmosfera (CARVALHO; SILVA, 2006), que é representado de maneira geral na Figura 1, onde é possível visualizar todos os possíveis fatores inerentes ao ciclo hidrológico.

Figura 1 – Ciclo hidrológico e seus fatores.



Fonte: Weather (2019), adaptado.

Correntemente a chuva apresenta uma grande variabilidade espacial, assim, durante um evento de chuva diferentes áreas podem receber índices pluviométricos [mm] distintos em uma janela de tempo, informação essa que é transformada para outras localizações em uma mesma região de interesse a partir da espacialização de chuvas considerando informações de geolocalização (COLLISCHONN; TASSI, 2008).

O elemento físico do sistema, bacia hidrográfica, é a área de captação natural ou modificada da água de precipitação que faz convergir os escoamentos para um único ponto de saída, o exutório. Por se tratar de um sistema aberto, nem toda a precipitação se torna escoamento no exutório ou fica armazenada na própria bacia (PAZ, 2004).

O tempo demandado para que a água precipitada em uma bacia hidrográfica se transforme em vazão e chegue ao exutório é conhecido como tempo de concentração, e este depende da morfologia da bacia.

Este processo é conhecido como processo chuva-vazão, transformando a chuva em escoamento superficial, sendo um dos mais complexos fenômenos hidrológicos que ocorrem em uma bacia hidrográfica, devido à variação temporal e espacial dos diversos elementos de cada evento hidrológico (SCHEIDT; ANGELICA; BRUNETTO, 2011).

Contudo, para fundamentação técnica em estudos envolvendo hidrologia, como as obras hidráulicas, é necessário o conhecimento das chuvas intensas que possam vir a ocorrer, sendo associada a outras informações de interesse, como a duração e frequência, na região de interesse (ALMEIDA, 2017).

A estimativa do tempo de retorno para ocorrência um determinado evento hidrológico é calculado baseado em dados históricos, e este é um parâmetro fundamental para gestão de recursos hídricos e desenvolvimento de obras hidráulicas, onde é estimado o intervalo de tempo para que uma dada chuva de intensidade e duração definidas seja igualada ou superada (TOMAZ, 2010).

Em obras de engenharia, a segurança e durabilidade da obra dependem do parâmetro tempo de retorno, que permite estimar a vazão de projeto para dimensionamentos de sistema de engenharia hidráulica, sendo baseado na necessidade e risco que estruturas hidráulicas possam resistir a enchentes. O valor de tempo de retorno em projetos depende da importância da obra e capacidade do bem feitor arcar com custos de instalação, manutenção e ou reparos (BRASIL, 2005).

A definição teórica de risco de falha de uma estrutura hidráulica é estimada na probabilidade de falha de ocorrer uma descarga de projeto com tempo de retorno − *TR* (em anos) dentro da vida útil da estrutura, fixada em *n* (anos), sendo estimada através da Equação 1, que é o risco hidrológico de falha (BRASIL, 2005; TOMAZ, 2010).

$$R = 1 - \left(1 - \frac{1}{TR}\right)^{n} \qquad (1)$$

Sabendo que a estimativa do risco hidrológico para falha é o inverso da probabilidade de falha, então o risco hidrológico indica que haverá maior probabilidade de falha para valores menores de tempo de recorrência e de anos. E quanto maior o *TR* e maior o *n*, menor o risco. Nestas circunstâncias, para determinada condição de *n*, deve-se prezar por um valor de *TR* razoável de modo que o projeto não tenha um elevado custo e as taxas de falhas possam ser mínimas (BRASIL, 2005).

Conforme a relação proposta na Equação 2, sabe-se que a probabilidade *P* de ocorrência de um evento hidrológico e a frequência *F* estão ligadas ao tempo de retorno estimado, contudo a frequência se torna mais evidente em séries históricas longas (TOMAZ, 2010).

$$P \approx F = \frac{1}{TR} \tag{2}$$

Os eventos de inundações e transbordamentos em sistemas de drenagem podem ser caracterizados como falha de processo ou evento natural. Embora os projetos de equipamentos de drenagem sejam realizados segundo parâmetros suficientemente seguros, não há viabilidade técnico econômica para um projeto indefectível à precipitação (BRASIL, 2005).

2.1.1 Tratamento de Dados Hidroclimatológicos

O estudo acerca da ocorrência de chuvas é primordial para pesquisas relacionadas à água, como modelagem hidrológica e modelagem empírica de processos afetados pelo mesmo, pois a precipitação é um fenômeno complexo, irregular e espacialmente descontínuo, frequentemente com acumulações nulas (DIRKS et al., 1998; PELLICONE et al., 2018).

Os dados diários de precipitação de chuva são um dos fatores básicos nos modelos hidrológicos e ambientais, portanto preencher possíveis dados faltantes nos dados diários de precipitação é uma questão essencial (HASAN; CROKE, 2013).

Dados pluviométricos faltantes podem ser preenchidos a partir do emprego de técnicas de interpolação, entretanto um dos problemas enfrentados nos estudos de padrões

espaciais de chuvas é a interpolação de dados de pluviômetros espaçados (DIRKS et al., 1998), pois a espacialização e fatores referentes a geomorfologia podem afetar os resultados (MIRÁS-AVALOS et al., 2007).

Além do preenchimento de dados, é comum também que métodos de interpolação sejam empregados para estimar parâmetros de equações de intensidade de chuva, como estudado e citado por Mello et al. (2003).

Um método geoestatístico determinístico amplamente utilizado para preenchimento de falhas é a interpolação pelo inverso da distância ponderada (*Inverse distance weighting - IDW*), Equação 3, que é baseado na hipótese de que o valor da precipitação em um ponto, não amostrado, pode ser estimado como a média ponderada pela distância a partir de valores de precipitação nos pontos com amostragem (PELLICONE et al., 2018).

$$Pm_i = \frac{\sum_{j=1}^{NP} \frac{P_j}{\left(d_{ij}\right)^b}}{\sum_{j=1}^{NP} \frac{1}{\left(d_{ij}\right)^b}} \tag{3}$$

no qual *NP* é o número de postos pluviométricos com dados disponíveis; P_j é a chuva observada no posto pluviométrico *j*; e d_{ij} é a distância euclideana, Equação 4, entre o centro da área de interesse e o posto pluviométrico.

$$d_{ij} = \left[\sum_{m=1}^{n} \left(x_{im} - y_{jm}\right)^2\right]^{1/2} \tag{4}$$

Considerando a distância entre postos, x_j e y_j são as coordenadas geográficas para cada ponto, Equação 5.

$$d_{ij} = \sqrt{\left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2} \tag{5}$$

Na Equação 3, quando o valor do expoente *b* é 2, o método de interpolação é conhecido como ponderado pelo inverso da distância ao quadrado (COLLISCHONN; TASSI, 2008).

Uma das grandes vantagens do IDW é a facilidade de implementação em ocasiões com disponibilidade de informação geográfica (SIG), pois o método é baseado na geometria da captação (BALL; LUK, 1998), e o principal parâmetro de influência crítica do IDW é a distância (CHEN; LIU, 2012).

Presente em diversos estudos, o método IDW foi utilizado satisfatoriamente para determinar a distribuição espacial de chuvas, como os exemplos da bacia do rio Upper Parramatta em Sydney/Austrália (BALL; LUK, 1998), para a ilha Norfolk (DIRKS et al., 1998), na região do Himalaia/Índia (BASISTHA; ARYA; GOEL, 2008), em Taiwan/China (CHEN; LIU, 2012), no estado da Bahia/Brasil para análise anual no período de 1981 à 2010 (DOURADO, 2013), na bacia do rio Xinxie na China (CHEN et al., 2017), na região da Calábria/Itália (PELLICONE et al., 2018).

2.2 SISTEMAS DE DRENAGEM

Os sistemas de drenagem pluvial são, geralmente, um conjunto de elementos de infraestrutura designados para coleta, transporte e retenção de água pluviais (CEMBRANO et al., 2004). Se situados em cidades ou povoamentos, são denominados sistemas de drenagem urbano. E, em caso de agrupado ao sistema de esgotamento sanitário, o sistema passa a ser denominado como sistema de esgotamento combinado.

A concepção de projeto e características do sistema de drenagem pluvial urbana depender da vazão de projeto a ser estimada em função do fenômeno chuva-vazão e a análise de risco, diferentemente das zonas ou sistemas industriais, a vazão de projeto engloba também características da indústria em questão (SCHMITT; THOMAS; ETTRICH, 2004). Contudo, um sistema de drenagem industrial pode ser projetado com foco em drenagem de efluentes industriais relacionados aos processos, sendo a chuva um agente externo que pode ou não contribuir com incremento de vazão.

Para atenuar a vazão de escoamento superficial oriunda do fenômeno chuva-vazão em determinados locais, medidas estruturais podem ser tomadas visando minorar os picos de cheia em bacias hidrográficas. Uma das principais técnicas empregadas é a criação dos tanques de retenção, popularmente conhecidos como “piscinão”, que são destinados a conter o excesso de chuva e proteger áreas à jusante (BASTOS, 2009).

Além do controle dos picos de vazão, cheias, os tanques de detenção de águas pluviais também são uma ferramenta ambiental útil contra a poluição das águas pluviais.

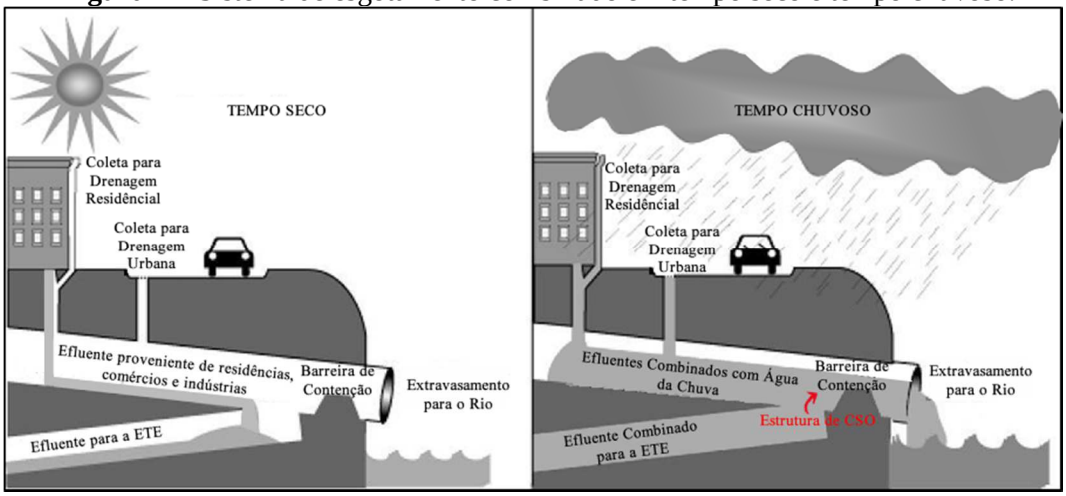
No entanto, em zonas urbanas e ou rurais, há a possibilidade de impacto negativo quanto a redução de vazões ecológicas para ambientes fluviais (TODESCHINI; PAPIRI; CIAPONI, 2012).

Em localidades onde o sistema de drenagem urbana é conjugado com o sistema de esgotamento sanitário, que acarreta em vazões maiores, estudos indicam que o desempenho do sistema melhora quando há aumento de vazão em função da chuva, mas a vazão incremental pode exceder a capacidade das estruturas hidráulicas do sistema e ocasionar o CSO (*Combined Sewer Overflow*), que é o transbordamento de águas residuárias (drenagem + efluente sanitário) para corpo hídrico, que implica em um disposição de efluentes possivelmente fora dos padrões ambientalmente aceitáveis (MARTINO et al., 2011).

O lançamento das águas residuárias sem o devido tratamento pode exceder os limites de poluição permitidos de acordo com as condições ambientais locais e os padrões governamentais. Para maior eficiência dessas estruturas hidráulicas, estruturas de CSO, costuma-se empregar tanques de águas pluviais e ou tanques/bacias de retenção, pois são muito eficazes para o controle da poluição devido ao escoamento de águas pluviais em áreas urbanas (MARTINO et al., 2011).

A estrutura de CSO pode ser visualizada na Figura 2, onde são apresentados os sistemas de drenagem urbana e esgotamento sanitário combinados em situações de tempo seco (sem ocorrência de chuvas) e tempo chuvoso.

Figura 2 – Sistema de esgotamento combinado em tempo seco e tempo chuvoso.



Fonte: EPA (2004), adaptado.

A ocorrência de eventos indesejáveis em sistemas hidráulicos podem ser classificadas como falhas, como a ocorrência de CSO em sistemas de drenagem urbana (THORNDAHL; SCHAARUP-JENSEN; JENSEN, 2008).

Estudos recentes de análise estatística de ocorrência de CSO foram realizados em Berlim na Alemanha (SANDOVAL et al., 2013), em Tóquio no Japão (YU et al., 2013, 2018), em uma região do norte do Reino Unido (MOUNCE et al., 2014), em Quebec no Canadá (MAILHOT; TALBOT; LAVALLÉE, 2015), em La Garriga na Espanha (MONTSERRAT et al., 2015), em uma área piloto no Reino Unido (ZHAO; BEACH; REZGUI, 2017), e Pau no sudoeste da França (BERSINGER et al., 2018).

Sandoval et al. (2013) abordaram a relação entre variáveis de precipitação pluviométrica, vazão e características de qualidade da água dos CSO, utilizando correlação canônica e regressão por mínimos quadrados parciais, e constataram que a intensidade, duração da chuva, e o tempo seco foram as principais variáveis para elaboração do modelo.

No estudo desenvolvido em Tóquio, foi realizado agrupamento de dados, análise de correlação por índice de similaridade e foi proposto um modelo de regressão empírico baseado na caracterização dos parâmetros de precipitação, em dados de simulação de escoamento superficial e do limiar de intensidade de precipitação (YU et al., 2013, 2018).

A similaridade entre os dois grupos categorizados de precipitação e transbordamento pode ser calculada baseada na Equação (6) referente ao Índice de Similaridade (IS) onde o $IS \geq 0.75$ indica que há similaridade entre os *clusters* (agrupamentos) construídos (YU et al., 2013).

$$IS = \frac{2N}{(N1 + N2)} \tag{6}$$

sendo *N* o número de eventos de precipitação categorizados no mesmo grupo tanto pelos grupos referente a precipitação quanto de transbordamentos; *N1* o número de eventos de precipitação no grupo de padrões de precipitação e *N2* o número de eventos de precipitação no grupo de comportamento de transbordamentos.

Mouce et al. (2014) utilizaram redes neurais artificiais para predizer o nível do efluente em uma estrutura de esgoto combinado, utilizando como uma das covariáveis a informação de intensidade de chuva obtida por imagem de radar, proporcionando uma

aplicação a tempo real onde foi possível estimar o nível com cinco passos à frente dos dados analisados, com apenas 5% de erro.

Em Quebec, Mailhot; Talbot e Lavallée (2015) utilizaram um modelo estatístico de função de probabilidade binomial para estimar a ocorrência de CSO em determinado dia, hora ou período sem ocorrência de CSO.

Em um sistema de drenagem urbana monitorado na cidade de La Garriga, Montserrat et al. (2015) elaboraram metodologia com o objetivo final fornecer informações de suporte à tomada de decisão, onde cada estrutura de CSO foi avaliada quanto à sua capacidade, desempenho e conformidade legal para fornecer apoio sobre a manutenção do sistema. Além disso foram elaboradas para cada estrutura uma árvore de decisão baseada em aprendizado de máquina onde a variável de saída foi a ocorrência ou não do transbordamento na estrutura de CSO.

Utilizando regressão LASSO, Zhao; Beach e Rezgui (2017) elaboraram um algoritmo para implementação de uma metodologia baseada na construção automatizada de modelos preditivos em tempo real de CSO usando dados de monitoramento de campo, proporcionando uma nova opção metodológica para estimativas de modelos chuva-vazão sem emprego direto de modelagem hidrológico-hidráulica.

Bersinger et al. (2018) desenvolveram estudo objetivado determinar estatisticamente os parâmetros que mais influenciam a concentração máxima de DQO (demanda química de oxigênio), o lançamento volumétrico de efluentes e o lançamento de carga de DQO durante um evento pluviométrico, para tanto utilizou árvore de regressão condicional.

O avanço e evolução técnico-científico neste ramo de pesquisa através do emprego de técnicas de predição é notório, pois com maior gama de dados históricos disponíveis, sendo possível mensurar vazões e ocorrência de eventos relacionados ao sistema em estudo, permitindo que situações indesejadas sejam gerenciadas de forma mais proativa.

É possível também, a partir da análise exploratória e modelagem do fenômeno de transbordamentos, identificar modos de falha e proporciona redução de custos com a modelagem de simulação convencional que envolve requisitos de informação de campo bem detalhadas (MOUNCE et al., 2014).

Entretanto, eventos de transbordamentos não estão restritos a sistemas de drenagem em meio urbano, podem ocorrer também nos sistemas em industrias, bem como qualquer sistema susceptível ao risco.

Em grandes industrias, podem ser adotados dois sistemas de drenagem independentes, o sistema de efluentes industriais, que é aquele projetado para os efluentes de processos industriais e as áreas diretamente ou potencialmente afetadas pelos processos, e o sistema de drenagem pluvial, que é projetado para as zonas urbanas dentro da área industrial. Tal fato ocorre porque Resolução CONAMA nº 20 de 1986, veda a possibilidade de diluição de efluentes industriais com fontes não poluídas.

2.3 APRENDIZADO DE MÁQUINA

A mineração de dados (*data mining*) é o processo de extração de informações implícitas, potencialmente úteis dos dados, de tal modo a elucidar regularidades ou padrões, podendo ser generalizados para fazer, por exemplo, boas predições sobre dados futuros (WITTEN et al., 2017).

Técnicas de aprendizado de máquina (*machine learning*) fornecem a base técnica da mineração de dados (WITTEN et al., 2017). James et al., (2013) preferem abordar este tema como aprendizado estatístico (*statistical learning*), referindo-se ao vasto conjunto de ferramentas estatísticas para entender os dados.

Essas ferramentas podem ser classificadas como supervisionadas (*supervised*), semi-supervisionadas (*semi-supervised*) ou não supervisionadas (*unsupervised*), diferindo basicamente na presença total, presença parcial ou ausência de rótulo aos dados.

Modelos de aprendizado supervisionado podem desempenhar um papel central no monitoramento de processos e no diagnóstico de falhas em industrias porque proporcionam resultados que servem como suporte para a tomada de decisão (CHEN; GE, 2019).

Contudo, o aprendizado não supervisionado auxilia a melhor compreensão acerca dos dados obtidos sobre os processos de tal modo a proporcionar melhor entendimento sobre as variáveis em análise, possibilidades de segmentação de dados, classificação baseado em análise exploratória e desenvolvimento de novas variáveis que estavam implícitas no processo (THOMAS; ZHU; ROMAGNOLI, 2018).

Para efetiva aplicação de técnicas de aprendizado de máquina é essencial que o conjunto de dados (*dataset*) seja devidamente tratado, podendo melhorar a qualidade geral dos dados para análise adicional e construção de modelos, que inclui quatro componentes: dados faltantes (*missing data*), detecção de valores aberrantes (*outliers*), remoção de ruído e alinhamento de tempo (XU et al., 2015).

Estratégias adequadas de monitoramento podem garantir uma boa qualidade da informação, entretanto sistemas multivariados com elevada quantidade de informação tornam complexa a tarefa de monitoramento. Em operação envolvendo águas residuárias a escolha por métodos ou sistemas de monitoramento que minimizem a quantidade de dados perdidos e que proporcionem capacidades adaptativas à operação devem ser priorizadas (ROSEN; RÖTTORP; JEPPSSON, 2003).

Além do ganho de informação inicial baseado nos resultados da análise na mineração de dados, é esperado que com o monitoramento contínuo e em larga escala dos processos seja possível melhorar as estratégias de controle de processos nas industrias (WEESE et al., 2016).

Contudo, uma etapa primordial antes de qualquer análise de mineração de dados é o pré-processamento de dados, pois nesta etapa deve-se verificar a possibilidade de derivação de novos atributos, variáveis, a partir das variáveis já existentes, bem como a realização de estratégia para agregar, aglutinar, informações que possam ser entendidas como repetitivas no conjunto de dados (FEELDERS; DANIELS; HOLSHEIMER, 2000).

Outro ponto importante relacionado ao pré-processamento é o de transformação dos dados, como a normalização, pois é possível melhorar a precisão e a eficiência dos algoritmos de mineração quando os dados estão em uma mesma escala, sem ponderação dimensional ou de magnitude (SHALABI; SHAABAN; KASASBEH, 2006).

A normalização dos dados é útil também para reduzir a assimetria dos atributos, reduzindo possíveis distorções em função das diferenças da escala dos dados dos atributos, como a *Z-score*, Equação 7 (KUHN; JOHNSON, 2013).

$$Z_i = \frac{x_i - \mu}{\sigma} \tag{7}$$

onde Z_i é o novo valor para o objeto em cada atributo; x_i é o valor original do objeto no atributo; μ e σ são respectivamente a média e o desvio padrão por atributo.

2.3.1 Aprendizado Não Supervisionado

Aprendizado de máquina não supervisionado é uma técnica que permite descobrir as relações existentes entre objetos de um conjunto de dados descrito por um conjunto de atributos/variáveis.

Em situações de ausência de rótulo os atributos para os dados, o trabalho a ser desenvolvido se torna mais desafiador, pois tende a ser mais subjetivo, e não há objetivo simples para a análise, como a previsão de uma resposta, sendo assim indicada como parte de uma análise de dados exploratória (JAMES et al., 2013).

Primordialmente, busca-se encontrar padrões, objetos com características similares, para vencer a dificuldade da ausência de rótulo, e os padrões podem ser identificados realizando agrupamento de objetos, *clustering*.

Dentre os principais métodos de agrupamento pode-se destacar os algoritmos com método hierárquico, estratégia de partição e estratégia baseada em densidade.

2.3.1.1 Agrupamento de objetos

O método de Ward é um método de agrupamento hierárquico aglomerativo, onde inicialmente são agrupados os objetos ou grupos mais semelhantes, onde a cada iteração de execução do algoritmo os grupos são agregados dois a dois até a totalidade de objetos em análise (PESSANHA et al., 2015).

O método de Ward (WARD, 1963) busca unir dois grupos cuja a fusão será à menor soma de quadrados dentro do grupos, proporcionando variação mínima dentro do grupo de tal modo a produzir grupos de tamanho igual e compactos (PIOT, 2014). A fusão de grupos a cada iteração é objetivada na minimização da perda de informação associada a cada grupo, que é quantificada no critério da soma dos quadrados dos erros (*error sum-of-squares* - ESS), Equação 8 (HÄRDLE; HLÁVKA; KLINKE, 2000).

$$ESS = \sum_{k=1}^{K} \sum_{x_i \in C_k} \left(x_{ij} - \bar{x}_{kj}\right)^2 \qquad (8)$$

onde $\bar{x}_{kj}$, Equação 9, é a média interna em cada grupo; x_{ij} o valor para o *i*-ésimo indivíduo no *j-grupo*; *k* é o número total de grupos em cada estágio/iteração; e n_j é o número de indivíduos no *j*-ésimo grupo.

$$\bar{x}_{ij} = \frac{1}{n_k} \sum_{x_i \in C_k} x_{ij} \tag{9}$$

O *K-means* é um algoritmo de agrupamentos de dados por partição que tem como objetivo encontrar partições no conjunto de dados de forma de forma iterativa a gerar *k* grupos distintos a partir do chute inicial para o valor de *k*, a partir da escolha aleatória dos vetores distintos que representam os centroides (SILVA; PERES; BOSCARIOLI, 2016).

A partição dos grupos pode ser interpretada a partir do pseudo código genérico apresentado a seguir para o algoritmo do *k-means* (PIECH, 2019):

Algoritmo: *k-means*	
1	Escolha de *k* grupos aleatoriamente, indica-se o uso das médias;
2	Cálculo dos *k* centroides baseado em Distância Euclideana;
3	Atribuição de x_i ao centroide k_i mais próximo;
4	Repetição baseado na Equação 10 e Equação 11 iterativamente para reposicionamento do centroide até a convergência e estabilidade dos agrupamentos

$$argmin_s = \sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|^2 = argmin_S \sum_{i=1}^{k} |k_i| Var k_i \tag{10}$$

$$\therefore \sum_{x \in S_i} \|x - \mu_i\|^2 = \sum_{x \neq y \in S_i} (x - \mu_i)(\mu_i - y) \tag{11}$$

onde o x_i pertence ao conjunto de observações $\{x_1, x_2, ..., x_n\}$; μ_i é a média de pontos em k_i; e o valor de *k* é menor que *n*, $k = \{k_1, k_2, ..., k_k\}$.

O algoritmo PAM (KAUFMAN; ROUSSEEUW, 1990) é similar ao *K-means*, entretanto os centroides, isto é, medóides, pertencem ao conjunto de dados. O PAM é mais robusto para *outliers* quando comparado ao *K-means*, sua partição pode ser interpretada a partir do pseudo código genérico apresentado a seguir para o algoritmo do PAM (KASSAMBARA, 2017).

Algoritmo: PAM	
1	Seleção de *k* para se tornar medóides;
2	Cálculo de similaridade baseado em Distância Euclideana;
3	Repetição análoga ao *K-means*.

A qualidade interna e estabilidade dos agrupamentos gerados pode ser verificada numericamente.

2.3.1.1.1 Métodos de validação de agrupamentos

Visto o grande números de grupos existentes, decidir qual método de agrupamento usar e qual o número ótimo de grupos é uma tarefa difícil para o pesquisador que conduz o trabalho. Os grupos gerados não devem apenas ter boas propriedades estatísticas (compactos, bem separados, conectados e estáveis), mas também devem apoiar a interpretação do fenômeno em análise (BROCK et al., 2008).

A validação do agrupamento pode ser feito baseado na validação externa, validação interna e estabilidade. As medidas de validação externa são baseadas em entropia e necessitam de informação conhecida a priori, como os rótulos. Portanto, na situação em que não há informações externas disponíveis, as medidas de validação interna são a única opção para validação do grupo (LIU et al., 2010).

Segundo Handl, Knowles e Kell (2005) a medida de conectividade é uma medida de validação interna que mensura a relação de similaridade a medida em que as observações são agrupadas no mesmo grupo que seus vizinhos mais próximos no espaço de dados (BROCK et al., 2008).

A conectividade é definida considerando um conjunto com N elementos com b variáveis em k grupos $\{k_1, ..., k_k\}$, com L é o número de vizinhos mais próximos de um objeto x_i, define como $nn_{i,j}$, j-ésimo vizinho mais próximo da observação i. Se $L=1$, então apenas o elemento mais próximo de x_i é considerado seu vizinho (BROCK et al., 2008; PESSANHA et al., 2015).

$$Conectividade = \sum_{i=1}^{N} \sum_{j=1}^{L} z_{i,j} \qquad (12)$$

onde $z_{i,j} = 0$, se x_i e $nn_{i,j}$ pertencem ao mesmo grupo. Caso x_i e $nn_{i,j}$ não pertencem ao mesmo grupo então $z_{i,j} = 1/j$. A Conectividade na partição ideal em k agrupamentos deve ser minimizada, tal que $Conectividade \in [0, \infty)$.

De acordo com Brock et al. (2008) e Pessanha et al. (2015) a silhueta de um objeto x_i permite avaliar se o mesmo foi bem classificado entre os k grupos possíveis, sendo definido pela Equação 13:

$$Silhueta_i = \frac{b_i - a_i}{max(b_i, a_i)} \tag{13}$$

onde a_i , Equação 14, é a média das distâncias entre x_i e os objetos classificadas no mesmo grupo k, , e b_i , Equação 15, é a média das distâncias entre x_i e os objetos no grupo vizinho mais próximo do grupo k.

$$a_i = \frac{1}{n\big(k(i)\big)} \sum_{j \in C(i)} dist(i,j) \tag{14}$$

$$b_i = min \sum_{j \in k_k} \frac{dist(i,j)}{n(k_k)} \tag{15}$$

sendo $k(i)$ o grupo que contém o objeto i, $dist(i, j)$ é a distância entre os objetos i e j, e $n(k)$ é a cardinalidade do grupo k. A largura da silhueta deve ser maximizada, tal que $Silhueta \in [-1,1]$.

Outra métrica de validação interna é o índice Dunn, Equação 16, que é definido pela razão da menor distância entre duas observações em grupos distintos pela maior distância entre dois grupos (BROCK et al., 2008; PESSANHA et al., 2015). É calculado como:

$$Dunn = \frac{\min\limits_{k_k, k_l \in k, k_k \neq k_l} \left(\min\limits_{i \in k_k, j \in k_l} dist(i,j) \right)}{\max\limits_{k_m \in k} diam(S_m)} \tag{16}$$

sendo $diam(k_m)$ a máxima distância entre os objetos no grupo k_m. O índice Dunn deve ser maximizado, tal que $Dunn \in [0, \infty)$.

A avaliação de estabilidade em grupos é mensurada através de amostragens repetitivas dos dados para gerar agrupamentos. Onde um algoritmo “bom” deverá identificar agrupamentos que não variam muito de um conjunto amostral para outro, apresentando estabilidade em relação à aleatorização de entrada (BEN-DAVID; PÁL; SIMON, 2007). Que significa dizer que as características dos grupos devem permanecer semelhantes para diferentes amostragens (HENNIG, 2007).

As principais métricas para análise de estabilidade são a proporção média de não-sobreposição (*Average proportion of non-overlap* - APN), a distância média (*Average distance* - AD), a distância média entre médias (*Average distance between means* - ADM) e a figura de mérito (*Figure of merit* - FOM), as quais apresentam bom funcionamento especialmente quando os dados em análise são altamente correlacionados (BROCK et al., 2008).

A proporção média de não-sobreposição (APN), Equação 17, mensura a proporção média de observações não colocadas no mesmo grupo na realização do agrupamento com base no conjunto de dados completos, e em grupo com base nos dados quando uma única variável é removida, sendo $APN \epsilon [0, 1]$ tal que os valores próximos de zero correspondem a resultados de agrupamento estáveis (PIHUR; BROCK; DATTA, 2009).

$$APN(C) = \frac{1}{MN} \sum_{i=1}^{N} \sum_{l=1}^{M} \left(1 - \frac{n(C^{1,l} \cap C^{1,0})}{n(C^{1,0})} \right) \tag{17}$$

onde $C^{i,0}$ representa o grupo que contém a observação *i* para os agrupamentos utilizando todas as variáveis; $C^{i,l}$ representa o grupo contendo a observação *i* para os agrupamentos realizados com a remoção da variável *l*; *N* é o número total de objetos; *M* é o número total de variáveis; e *n* é a cardinalidade de cada conjunto. Tais variáveis também estão presentes na formulação da distância média (AD).

A distância média (AD), Equação 18, é definido como a distância média entre os objetos colocados em uma mesmo grupo na realização do agrupamento com base no conjunto de dados completos, e no grupo com base nos dados quando uma única variável é removida, sendo $AD \, \epsilon [0, \infty)$ tal que os valores próximos de zero correspondem a resultados de agrupamento estáveis (PIHUR; BROCK; DATTA, 2009).

$$AD(C) = \frac{1}{MN}\sum_{i=1}^{N}\sum_{l=1}^{M}\left(\frac{1}{n(C^{1,0})n(C^{i,l})}\right)\left[\sum_{i\in C^{i,0}, j\in C^{i,l}} dist(i,j)\right] \quad (18)$$

Distância média entre médias (ADM), Equação 19, é definido como a distância média entre os centros dos grupos para objetos colocadas no mesmo grupo por agrupamento com base no conjunto de dados completo, e quando o agrupamento é baseado na remoção quando uma única variável é removida, sendo $ADM \epsilon [0, \infty)$ tal que os valores próximos de zero correspondem a resultados de agrupamento estáveis (PIHUR; BROCK; DATTA, 2009).

$$ADM(C) = \frac{1}{MN}\sum_{i=1}^{N}\sum_{l=1}^{M} dist\left(\bar{x}_{C^{i,l}}, \bar{x}_{C^{i,0}}\right) \quad (19)$$

onde $\bar{x}_{C^{i,0}}$ é a distância média dos objetos no grupo que contém objeto *i* quando o agrupamento é realizado no conjunto de dados completos; $\bar{x}_{C^{i,l}}$ é a distância média dos objetos que contém o objeto *i* quando o agrupamento é realizado com a remoção da variáveis *l*.

A figura de mérito (FOM), Equação 20, mensura a variância média intra-grupo dos objetos utilizando a variável removida (*l*), sendo o resultado a média de todas as colunas removidas onde $FOM \in [0, \infty)$ tal que os valores próximos de zero correspondem a resultados de agrupamento estáveis (PIHUR; BROCK; DATTA, 2009).

$$FOM(l, C) = \sqrt{\frac{N}{N-K}} \times \sqrt{\frac{1}{N}\sum_{k=1}^{K}\sum_{i\in C_k(l)} dist\left(x_{i,l}, \bar{x}_{C_k(l)}\right)} \quad (20)$$

onde $x_{i,l}$ é o valor por objeto *i* na *l*-ésimo coluna no grupo $C_k(l)$; $\bar{x}_{C_k(l)}$ é a média do grupo $\bar{x}_{C_k(l)}$; *K* é o total de grupos; *N* é o número total de objetos.

2.3.1.2 Change-point

Os métodos de aprendizado de máquina não supervisionados não estão contidos apenas em abordagens relacionadas a objetos, é possível também, a partir dos diversos

algoritmos de agrupamento, realizar agrupamento de séries temporais uni ou multivariadas.

O comportamento acerca das séries temporais pertinentes a sistemas ambientais e seus efeitos, como o processo chuva-vazão e alteração no nível de bacias de contenção, é complexo, mas também pode ser estudado através da investigação de uma abordagem quantitativa e qualitativa relacionada a séries temporais. Nesta temática, também é oportuno e apropriado o estudo de detecção de pontos associados a mudanças bruscas (*change-points*) quanto ao comportamento das séries temporais.

Os *change-points* são os pontos identificados onde é possível segmentar a(s) série(s) temporais a fim de obter análise de tendências e estatística descritiva para cada segmento, e eles podem ocorrer de acordo com os parâmetros estatísticos média e ou variância da série em análise (COSTA; GONÇALVES; TEIXEIRA, 2016). Em situação para verificação para média e variância, deve-se seguir o teste de hipótese apresentado na Equação 21 e Equação 22.

$$H_o: \mu_1 = \cdots = \mu_n = \mu \wedge \sigma_1^2 = \cdots = \sigma_n^2 = \sigma^2 \tag{21}$$

$$H_1: \mu_1 = \cdots = \mu_k \neq \mu_{k+1} = \cdots = \mu_n \wedge \sigma_1^2 = \cdots = \sigma_k^2 \neq \sigma_{k+1}^2 = \cdots = \sigma_n^2 \tag{22}$$

A identificação dos pontos de mudança podem ser realizadas através de estimativas de informação a priori, como o critério de informação de Schwarz (*SIC*), Equação 23, que é baseado na máxima verossimilhança de um dado modelo penalizado pelo número de parâmetros estimados no modelo (TEIXEIRA, 2012). O modelo, segmento, que minimiza o *SIC* é considerado o mais adequado em comparação com o modelo sem ponto de mudança.

$$SIC_j = -2ln\left(L\left(\widehat{\Theta}_j\right)\right) + p_j ln(n), \qquad j = 1, 2, \ldots, M \tag{23}$$

onde *n* é o tamanho da amostra; $L\left(\widehat{\Theta}_j\right)$ é a máxima verossimilhança para o modelo *j*; $\widehat{\Theta}_j$ é uma estimativa de Θ_j; p_j é o número de parâmetros estimados no modelo *j*.

Através da análise de *change-point* é possível segmentar a série temporal, como séries hidroclimáticas, em um ou mais pontos, possibilitando a avaliação de mudanças em diferentes instantes, onde cada segmento da série pode ser interpretado como um

grupo (BARRETO et al., 2017; BEAULIEU; CHEN; SARMIENTO, 2012; WANG; KILLICK; FU, 2014).

Da interpretação sobre séries temporais, uma sequência ordenada de dados, $y_{1:n}$ é igual a (y_1, ..., y_n), tal que um *change-point* significativo ocorre dentro deste conjunto quando existe um tempo $\tau \in \{1, ..., n-1\}$, na qual as propriedades estatísticas dos segmentos de série (y_1, ..., y_τ) e ($y_{\tau+1}$, ..., y_n) são diferentes por média e ou variância (KILLICK; ECKLEY, 2014).

A inferência estatística a partir do *change-point* para detecção de múltiplos pontos de mudança pode ser realizada por meio de algoritmos computacionais, como algoritmo de segmentação binária (*binary segmentation*), *segment neighborhood* e PELT (*Pruned Exact Linear Time*) (COSTA; GONÇALVES; TEIXEIRA, 2016; KILLICK; ECKLEY, 2014).

O algoritmo de segmentação binária pode ser interpretado a partir do pseudo código genérico apresentado a seguir (TEIXEIRA, 2012).

Algoritmo: Segmentação Binária	
1	A série univariada é analisada;
2	Detecta-se uma única mudança considerando a sequência de observações completa; Se não existir nenhum *change-point* H_0 é aceita; Se existir um *change-point*, então divide-se a sequência original de observações em duas subsequências.
3	Para cada subsequência, segmento, reinicia-se o procedimento de teste de *change-point*, até não ser detectado nenhum *change-point* nos segmentos criados.

2.3.2 Aprendizado Supervisionado

Quando empregado técnica de aprendizagem supervisionada, o objetivo principal, normalmente, é prever o valor de uma medida ou classe com base no entrada (*input*) de informações, onde a presença previamente conhecida das respostas referente a variável dependente orienta o processo de aprendizagem.

Os modelos supervisionados tem sido amplamente utilizados para várias aplicações nas indústrias (GE et al., 2017). Um dos métodos supervisionados é o *KNN* (*K-nearest neighbors*), que é um método não-paramétrico (ausência de parâmetros estatísticos relacionados a distribuição de dados) baseado na votação dos vizinhos mais

próximos de acordo com métrica de distância para similaridade (SILVA; PERES; BOSCARIOLI, 2016).

Outro algoritmo não-paramétrico é o de Árvore de Decisão (AD), que é baseado em método de árvores, e são simples e úteis para interpretação para modelos de regressão ou classificação e levam vantagem frente a métodos paramétricos, principalmente pela flexibilização dos pressupostos estatísticos (JAMES et al., 2013).

Contudo, para a construção de um modelo preditivo de classificação, é ideal que os dados estejam balanceados, ou seja, a probabilidade de ocorrência de um evento deve ser igual ou próxima da probabilidade da não ocorrência do evento em interesse. A construção de modelos preditivos que são treinados com conjuntos de dados desbalanceados de classe são altamente suscetíveis a produzir modelos de previsão imprecisos, entretanto é possível utilizar técnicas de reamostragem (*resampling*) para contornar esta questão a partir dos dados observados, como *under-sampling*, *over-sampling* e ROSE (*Random Over Sampling Examples*) que é baseada em *bootstrap* suavizado (TANTITHAMTHAVORN; HASSAN; MATSUMOTO, 2018).

2.3.2.1 K-nearest neighbors - KNN

A abordagem preditiva a partir dos *K* vizinhos mais próximos, *K-nearest neighbors* (KNN), possibilita a predição de um novo objeto a partir de informação acumulada, em forma de dissimilaridade, de outros objetos mais próximos do conjunto de treinamento. Para regressão, o algoritmo KNN identifica os *k*-vizinhos mais próximos do objeto no espaço preditivo, e resposta prevista para o objeto é, de modo clássico, a média das respostas dos vizinhos. Em caso de classificação, é realizada votação para verificação da classe predominante na densidade do objeto de interesse (SILVA; PERES; BOSCARIOLI, 2016).

A categorização dos objetos após classificação é baseada no classificador de Bayes, onde estima-se a distribuição condicional da resposta *Y* dado a característica *X*, onde o objeto é classificado para a classe com maior probabilidade estimada, maior quantidade de votos (JAMES et al., 2013).

Da funcionalidade do KNN, usa-se os objetos no conjunto de treinamento *T* mais próximo no espaço de entrada para *x* para estimar uma resposta $\hat{Y}$, Equação 24. Onde o

k-vizinho mais próximo para $\hat{Y}$ é definido como (HASTIE; TIBSHIRANI; FRIEDMAN, 2009; JAMES et al., 2013):

$$\hat{Y} = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i \tag{24}$$

onde $N_k(x)$ onde é a vizinhança de x definida pelos k objetos mais próximos a x_i na amostra de treinamento.

O método KNN estima a distâncias entre objetos a partir da métrica de distância para similaridade, podendo ser a Distância Euclideana (KUHN; JOHNSON, 2013).

O número de vizinhos mais próximos k geralmente é escolhido empiricamente e pode variar a cada problema ou conjunto de dados, a escolha do hiperparâmetro k ótimo é baseado no modelo que proporciona melhor desempenho (precisão), podendo ser avaliado na etapa de validação cruzada, e uma das possíveis abordagens para estimativa do número ótimo de vizinhos mais próximos (k) compreende analisar o intervalo de $k = 1$ à $k = \sqrt{n}$ tal que $k \in N$, (HASSANAT; ABBADI; ALTARAWNEH, 2014; ZHANG et al., 2017).

2.3.2.2 Métodos baseados em árvores

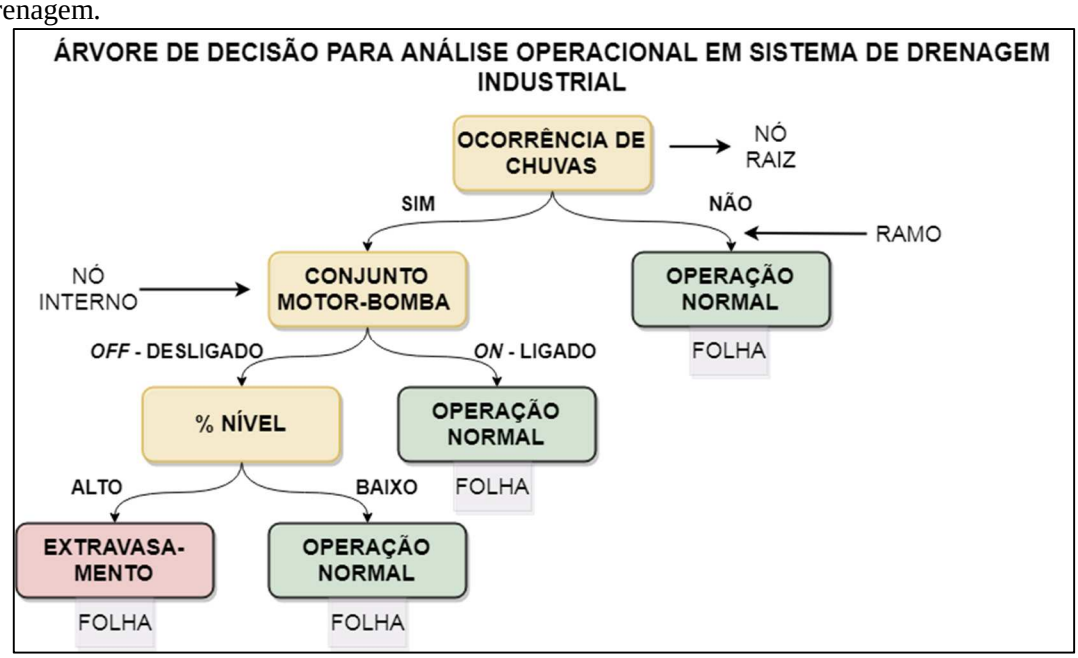
Dentre os métodos baseados em árvores, as Árvores de Decisão são modelos estatísticos que utilizam treinamento supervisionado para a classificação e previsão de dados, e deles pode-se extrair regras do tipo “se-então” que são facilmente compreendidas (SILVA, 2005). Essas regras de extração ou segmentação do espaço do preditor em um número de regiões simples (JAMES et al., 2013).

A Árvore de Decisão tem como estrutura básica o nó, ramo e folha. Os nós dizem respeito aos atributos, os ramos são as possíveis classes por atributo; cada nó folha representa a decisão/predição sugerida pelo modelo de classificação (SILVA; PERES; BOSCARIOLI, 2016), como pode ser visto em um caso hipotético na Figura 3.

Na Figura 3, é exemplificada uma Árvore de Decisão para verificação hipotética se um sistema de drenagem está em operação normal ou em transbordamento. O nó raiz, o primeiro nó interno, é a ocorrência de chuvas, e a partir deste é possível determinar se o sistema está em operação normal ou pode estar ocorrendo transbordamento. Em caso

de ocorrência de chuvas, outros nós podem ser analisados para determinação do estado operacional do sistema, que é verificado através do nó folha.

Figura 3 – Árvore de decisão hipotética para avaliação operacional em sistema de drenagem.



Fonte: O Autor, 2019.

Para a construção da Árvore de Decisão é empregado a abordagem *top-down* (de cima para baixo), do topo da árvore (nó raiz) depois divide-se (*split*) sucessivamente o espaço do preditor onde a cada divisão é indicada por dois novos ramos mais abaixo na árvore, que é conhecida como divisão binária recursiva (JAMES et al., 2013).

Esta abordagem de construção da árvore é gananciosa pois em cada etapa do processo de construção da árvore, a melhor divisão é feita sem verificar a possibilidade de melhoria em etapa (nó) futura (JAMES et al., 2013).

Para predição de classes, realiza-se a divisão binária recursiva, é selecionado o preditor X_n e o ponto de corte para o *split*, de forma a dividir o espaço preditivo em duas regiões R_n, ramos, $R_1(n,s) = \{X|X_n < s\}$ e $R_2(n,s) = \{X|X_n \geq s\}$ de modo a proporcionar a maior redução possível da taxa de erro de classificação (*classification error rate*) (JAMES et al., 2013).

A taxa de erro de classificação, Equação 25, considerando uma classificação binomial $y_i \in \{0,1\}$ e uma amostra de treinamento $\mathcal{D}_n = \{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$ onde x_i é o preditor, é estimada (KIM, 2009).

$$\epsilon_n = E\left[I\left(Y \neq R_n(X)\right)|\mathcal{D}_n\right] \quad (25)$$

onde $R_n(X) = R(\mathcal{D}_n, X)$ é um classificador construído baseado na amostra de treinamento $\mathcal{D}_n$. Sendo assim taxa de erro condicional é a probabilidade condicional de erro de classificação dada a amostra de treinamento.

As regiões particionadas, podem ser interpretadas como grupos menores e mais homogêneos, são mais puros, que contêm uma proporção maior de uma classe em cada nó posterior (KUHN; JOHNSON, 2013).

No entanto focar na minimização do erro de classificação pode não proporcionar o particionamento ótimo dos dados, sendo assim necessário implementação de métrica para buscar o particionamento com maior pureza, como o índice Gini, sendo definido pela Equação 26 (KUHN; JOHNSON, 2013).

$$Gini = p_1(1 - p_1) + p_2(1 - p_2) \quad (26)$$

onde p_1 e p_2 são as probabilidades ou frequências relativas para predição dicotômica classe 1 e classe 2, respectivamente. Como esse é um problema de duas classes, a soma das probabilidade p_1 e p_2 totaliza o valor de 1, portanto, Equação 26 é equivalente a Equação 27.

$$Gini = 2p_1p_2 \quad (27)$$

Diferente da taxa de erro de classificação, o índice de Gini é minimizado quando uma das probabilidades de classe é direcionada para zero, ou seja, o nó é puro em relação a uma das classes, e é maximizado quando p_1 é igual ou tem valor muito próximo a p_2, o caso em que o nó apresenta menor pureza (KUHN; JOHNSON, 2013).

Quando um conjunto de dados de duas classes é dividido em dois subconjuntos com base num ponto de divisão potencial de um atributo, é elaborado tabela de contingência, Tabela 1, para determinação dos valores do índice de Gini a priori, Equação 28, e pós *split*, Equação 29 (HARVEY; MCBEAN, 2014).

O melhor preditor de entrada é encontrado usando pesquisa exaustiva sobre os preditores a partir das amostras de conjunto de treinamento, sendo aquele que proporciona

menor impureza, o qual fornece o menor índice de Gini após o *split* (BREIMAN et al., 1984).

O algoritmo Árvore de Decisão tende a obter melhores resultados que métodos elementares como o KNN e resultados inferiores frente a regressão RIDGE-LASSO, situação que é contornada quando utilizado um modelo *ensemble*, como a *Random Forest* (JAMES et al., 2013).

Tabela 1 - Tabela de contingência a ser obtida após uma possível divisão feita na árvore.

	Classe 1	Classe 2	
> *Split*	n_{11}	n_{12}	n_{+1}
≤ *Split*	n_{21}	n_{22}	n_{+2}
	n_{1+}	n_{2+}	n

Fonte: Kuhn e Johnson (2013).

$$Gini\ (a\ priori\ do\ split) = 2\left(\frac{n_{1+}}{n}\right)\left(\frac{n_{2+}}{n}\right) \tag{28}$$

$$Gini\ (pós\ split) = 2\left[\left(\frac{n_{11}}{n}\right)\left(\frac{n_{12}}{n_{+1}}\right) + \left(\frac{n_{21}}{n}\right)\left(\frac{n_{22}}{n_{+2}}\right)\right] \tag{29}$$

Os métodos *ensemble* são algoritmos de aprendizado que possibilitam a construção de diversos classificadores individuais para posterior predição de classificação que serão obtidas através da votação das decisões dos classificadores individuais no conjunto (DIETTERICH, 2000a, 2000b).

Algoritmos como o *Random Forest*, que consiste na construção de Árvores de Decisão com aplicação de *bagging* para amostragem aleatória dos atributos e de objetos, é uma ferramenta eficaz para predição e robusta para *outliers* e ruído, bem como implementação em fenômenos com comportamento não-linear (BREIMAN, 2001). No qual *bagging* é uma técnica *bootstrap*, procedimento estatístico de amostragem com reposição, de atributos e objetos para obter várias previsões e médias (ou outras formas) dos resultados (WITTEN et al., 2017).

A implementação de *bagging* ou *bootstrap* com Árvore de Decisão CART (*Classification and Regression Trees*), *Random Forest*, é uma técnica para reduzir a variância de uma função para estimativa de previsão, onde, quando realizado para classificação, um conjunto de Árvores de Decisão aleatórias contabiliza baseado em cada árvore a votação para classe prevista (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Devido a possibilidade de tratar pontualmente de cada variável, a cada nó, as árvores tem capacidade para mapear padrões de interações complexas nos dados e, se cultivadas em níveis suficientemente profundos, apresentam viés relativamente baixo (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Para desenvolvimento da *Random Forest*, o principal hiperparâmetro de ajuste é o número de preditores selecionados aleatoriamente, *mtry* ou simplesmente *m*, para escolher em cada divisão, *split*. E, como configuração do algoritmo da *Random Forest*, ao menos 1 000 árvores (*ntree* = 1 000) devem ser utilizadas, pois sabe-se que as florestas aleatórias são protegidas contra o *overfitting*, logo, o modelo não será afetado negativamente se um grande número de árvores for construído para a floresta (KUHN; JOHNSON, 2013).

Quando realizada classificação, o algoritmo da *Random Forest* pode ser interpretado a partir do pseudo código genérico apresentado a seguir (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Algoritmo: *Random Forest*	
1	Para $b=1$ à B; Realiza-se uma amostragem bootstrap de Z^b com tamanho N a partir dos dados de treinamento; Desenvolve-se T_B árvores de decisão aleatórias (construção de modelo) para os dados do *bootstrap*, repetindo recursivamente as etapas a seguir para cada nó terminal da árvores, até que n_{min} (tamanho de nó mínimo) seja atingido; Seleciona-se m (*mtry*) variáveis aleatoriamente a partir das p variáveis preditoras; Escolhe-se a melhor variável/ponto de divisão entre as m através da minimização do índice Gini; Divide-se o nó (parental) em dois nós-filhos.
2	*Output* do conjunto *ensemble* de árvores $\{T_b\}_1^B$;
3	Realizar predição; Classificação: Seja $\hat{C}_b(x)$ a predição de classe da *b*th árvore da *Random Forest*. Então $\hat{C}_{rf}^B(x) = voto\ majoritário\ em \{\hat{C}_b(x)\}_1^B$.

onde Z^b é o conjunto de dados para cada realização de *bootstrap*; N é o tamanho de objetos total no conjunto de dados; p é a quantidade total de variáveis independentes no conjunto de dados.

O processo de seleção de variáveis aleatórias *m*, onde $m \leq p$, no *input* de cada árvores proporciona a redução da variância e correlação entre as árvores. Normalmente o

melhor valor de *m* é $\sqrt{p}$, ou valores muito baixo próximo a 1 (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

A *Random Forest* computa um voto para a classificação de uma nova amostra, e a proporção de votos em cada classe do conjunto é o vetor de probabilidade previsto (KUHN; JOHNSON, 2013).

Contudo, ao realizar o *bootstrap* espera-se que 2/3 dos dados sejam utilizados durante a criação do modelo *Random Forest*, logo 1/3 da informação é negligenciada, sendo esta conhecida como *out-of-bag* (OOB) (JAMES et al., 2013).

O erro OOB resultante é geralmente utilizado como uma estimativa válida do erro de teste para o modelo construído na *Random Forest* (JAMES et al., 2013).

Através do índice de impureza Gini é estimado também, de maneira ponderada, o atributo de maior importância para a coleção de árvores do *Random Forest* (KUHN; JOHNSON, 2013).

2.3.2.3 Métodos de amostragem

Em aplicações práticas de problemas de classificação, é muito comum que uma classe seja muito mais prevalente do que as outras, e isto implica também no desbalanceamento dos resultados da predição.

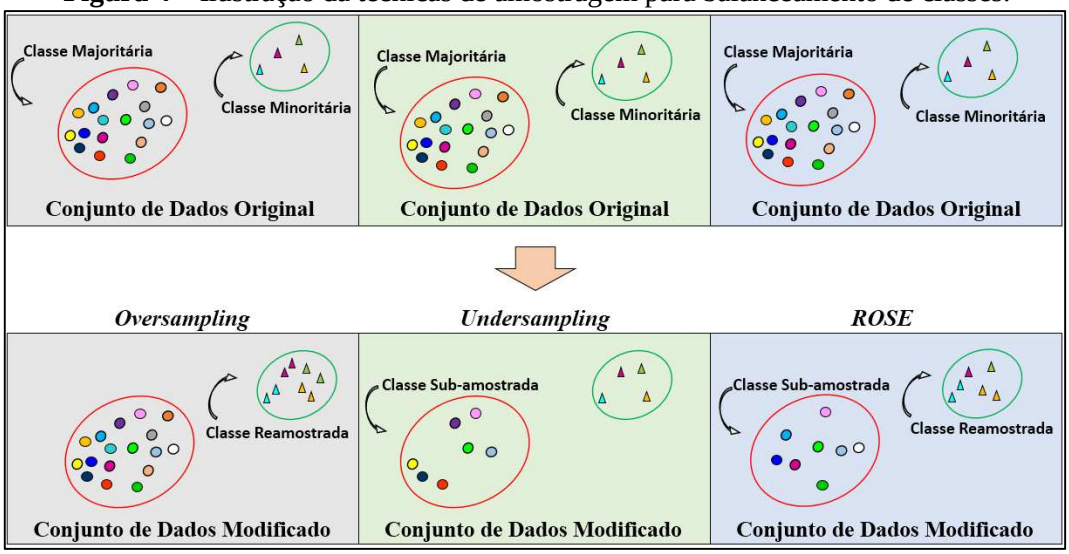
O problema de aprendizado desequilibrado resulta em redução de desempenho de algoritmos de aprendizado, pois eles pode resultar em elevado viés (*bias*), tendem a favorecer o grupo majoritário, e aprender com esses dados requer novos entendimentos, princípios, algoritmos e ou ferramentas para transformar os dados brutos de forma eficiente em representação de informação (HE; GARCIA, 2009; KRAWCZYK, 2016).

Para sanar o viés do desbalanceamento para predição existem muitas técnicas de rebalanceamento de classes, entretanto para previsão de um evento pode-se analisar a família de técnicas de amostragem, *undersampling* e *oversampling*, contudo técnicas de reamostragem baseadas *bootstrap* tendem a produzir estimativas mais precisas e confiáveis (TANTITHAMTHAVORN; HASSAN; MATSUMOTO, 2018).

As abordagens de *oversampling* são usadas para aumentar as amostras de dados na classe minoritária, enquanto a abordagem *undersampling* são usadas para reduzir as amostras de dados na classe majoritária, e o ROSE, desenvolvido por Menardi e Torelli

(2014), combina *undersampling* e *oversampling*, como pode ser visto na Figura 4 (LIN et al., 2017; TANTITHAMTHAVORN; HASSAN; MATSUMOTO, 2018).

Figura 4 – Ilustração da técnicas de amostragem para balanceamento de classes.



Fonte: O Autor, 2019.

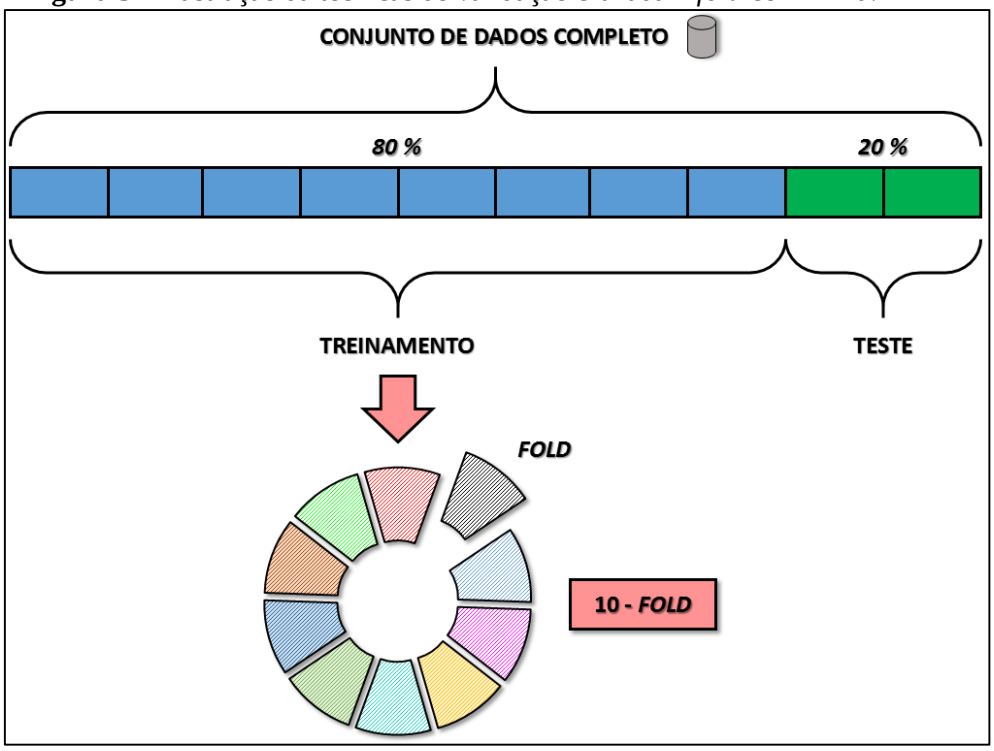
2.3.2.4 Validação de Modelos

Na aprendizagem supervisionada, os algoritmos de aprendizado são comparados de acordo com seu desempenho médio, que é formalmente definido pelo valor esperado do erro de predição ou quanto à classificação em relação aos conjuntos de teste, tal que se a quantidade de dados for grande o suficiente, o erro de predição pode ser estimado pelo erro médio em um conjunto de testes de espera, validação (BENGIO; GRANDVALET, 2003).

A técnica estatística de validação cruzada (*cross-validation*) permite estimar a os erros e habilidades (*skills*) dos modelos preditivos construídos. Com emprego do método *k-fold*, o procedimento da validação cruzada tem um único parâmetro chamado *k*, que diz respeito ao número de subgrupos para os quais uma determinada amostra de dados de treinamento deve ser dividida (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

A validação cruzada *k-fold* quando um valor específico para *k* é escolhido, como *k*=10, torna-se uma validação cruzada de 10 vezes, ou *ten folds*, como ilustrado na Figura 5 para um exemplo que 80% do conjunto de dados é utilizado para treinamento e 20% para teste, 72-8-20 (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Figura 5 – Ilustração da técnicas de validação cruzada *k-fold* com *k=10*.



Fonte: O Autor, 2019.

2.3.2.5 Avaliação de performance de modelos

Em modelos preditivos para classificação o *output* final é uma classe, categórica. Os modelos de classificação podem ser binomial ou multiclasses. Por exemplo, em um modelo binomial as respostas podem ser Sim ou Não, em um modelo com três classes poderia ser Sim, Não ou Talvez.

A verificação das probabilidades de classificação é um método eficaz de comunicar os resultados do modelo, pois os modelos construídos fornecem como resultado um valor de numérico contínuo de previsão, que geralmente é apresentado na forma de probabilidade ocorrência ou percentual de votos (ou seja, os valores previstos de participação na classe para qualquer amostra individual estão entre 0 e 1 e soma 1) (KUHN; JOHNSON, 2013).

Esta trama de análise baseada em probabilidade é derivada do classificador de Bayes, onde uma observação à classe para a qual a probabilidade posterior $p_k(X)$ é maior.

No caso de duas classes, isso equivale a atribuir uma observação à classe padrão se Pr (padrão = Sim | X = x) > 0,5 (JAMES et al., 2013).

Previsões para eventos binário simples (dicotômicas) representam o tipo mais simples de previsão e situação de tomada de decisão, onde os 2×2 resultados possíveis (contingências) para um evento são mostrados em uma matriz de confusão, também chamada de tabela de contingência, como a Tabela 2 (JOLIFFE, I; STEPHENSON, 2012).

Tabela 2 - Matriz de confusão a ser obtida após teste do modelo preditivo.

Evento Predito	Evento Observado	
	Transbordamento	Operação Normal
Transbordamento	Verdadeiro Positivo (VP)	Falso Positivo (FP)
Operação Normal	Falso Negativo (FN)	Verdadeiro Negativo (VN)

Fonte: O Autor, 2019.

Para uma sequência de previsões binárias, busca-se medidas de desempenho que possam ser formuladas em função dos verdadeiros positivos (VP) que representa número de acertos, os falsos positivos (FP) que são os alarmes falsos, os falsos negativos (FN) que são os erros, e os verdadeiros negativos (VN) que são as rejeições corretas (JOLIFFE, I; STEPHENSON, 2012).

A partir da matriz de confusão é possível derivar métricas estatísticas para avaliar o desempenho dos modelos construídos para classificação, algumas delas são acurácia, sensibilidade, especificidade, *Kappa* (COHEN, 1960), *F1* e ORSS. (HOSSIN; SULAIMAN, 2015). Essas métricas, exceto a *Kappa*, retornam resultado $\in [0, 1]$, sendo que quanto mais próximo a 1, melhor o resultado. A estatística *Kappa* retorna valores $\in [-1, 1]$.

A acurácia, Equação (30), é uma medida que corresponde a taxa de acertos, verdadeiro positivo e verdadeiro negativo, em todos os objetos.

$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN} = \frac{VP + VN}{n} \tag{30}$$

A sensibilidade, Equação 31, é a capacidade de determinar os eventos a ser predito corretamente, transbordamento, sendo uma proporção de verdadeiros positivos do evento de interesse.

$$Sensibilidade = \frac{VP}{VP + FN} \tag{31}$$

A especificidade, Equação (32), é a capacidade de determinar corretamente os eventos habituais, operação norma, sendo uma proporção de verdadeiros negativos em dos eventos habituais.

$$Especificidade = \frac{VN}{VN + FP} \tag{32}$$

O índice *Kappa*, Equação (33), compara a acurácia observada (p_0) com uma precisão esperada (p_e), que está diretamente relacionada ao número de instâncias de cada classe em análise juntamente com o *score* das predições assertivas (COHEN, 1960)

$$Kappa = \frac{p_o - p_e}{1 - p_e} \tag{33}$$

$$\therefore p_0 = Acurácia \wedge p_e = \left(\frac{VP + FP}{n} \times \frac{VP + FN}{n}\right) + \left(\frac{FN + VN}{n} \times \frac{FP + VN}{n}\right)$$

A métrica *F1*, Equação (34), proporciona uma resposta equilibrada quanto tratando principalmente de uma ponderação da sensibilidade.

$$F1 = 2 \times \frac{\frac{VP}{VP + FP} \times Sensibilidade}{\frac{VP}{VP + FP} + Sensibilidade} \tag{34}$$

Em situações de predições desbalanceadas métricas como a estatística *Kappa* e *F1* proporcionam uma ganho de informação para melhor compreensão acerca da predição em análise, pois valores elevados de sensibilidade e especificidade nem sempre traduzem um bom modelo (GOUTTE; GAUSSIER, 2005).

Além dessas métricas clássicas, é possível implementar métricas que de classificação para eventos meteorológicos, como a ORSS *(Odds Ratio Skill Score* – índice de taxa de probabilidade), Equação (35), aplicado por Mailhot; Talbot e Lavallée, (2015) para predição de transbordamento em sistema de drenagem urbano.

$$ORSS = \frac{(VP \times VN - FP \times FN)}{(VP \times VN + FP \times FN)} \tag{35}$$

3 METODOLOGIA

No presente trabalho foi realizado pesquisa aplicada, onde técnicas de mineração de dados para análise de dados e construção de modelos preditivos foram aplicadas, no qual o objeto de estudo são a operação do sistema de coleta e retenção de efluentes industriais em uma refinaria de petróleo, e a ocorrência de falhas, transbordamentos, desse sistema.

Para tal foram realizadas abordagens quantitativas para análise exploratória e construção de modelos preditivos, e qualitativas para melhor discussão do fenômenos em análise, dos atributos e da possibilidade de emprego dos produto gerado como elemento a suporte a tomada de decisão.

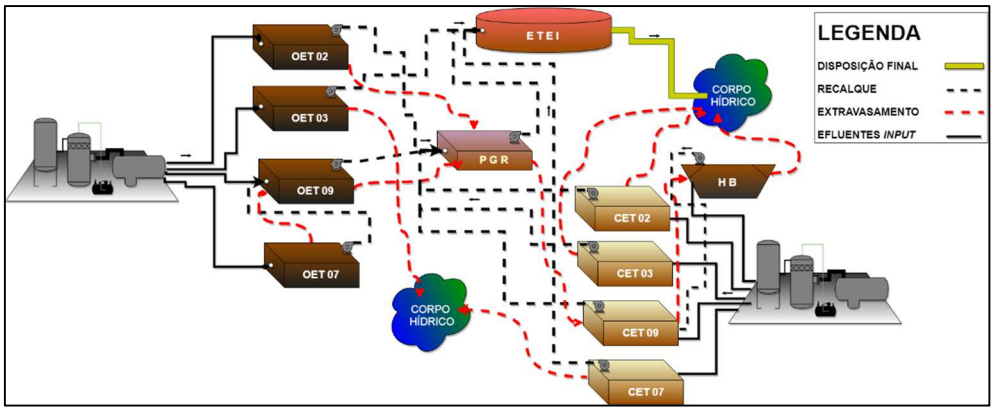
Foi realizado também pesquisa bibliográfica para levantamento de material técnico-acadêmico publicado sobre o fenômeno em análise e possíveis ferramentas a serem empregadas, de modo a proporcionar maior conhecimento na escolha das análises desenvolvidas.

Informações sobre o sistema de drenagem de efluentes industriais analisado, os dados obtidos e os métodos utilizados para modelagem do fenômeno de transbordamento estão descritos nas seções posteriores.

4 OBJETO DE ESTUDO

O sistema de efluentes industrias escolhido para o estudo de caso é formado por oito bacias de contenção de efluentes industriais de uma refinaria localizada próximo a um estuário, como pode ser visualizado de forma simplificada e ilustrativa na Figura 6.

Figura 6 - Representação simplificada do sistema de drenagem industrial no estudo de caso.



Fonte: O Autor, 2019.

Calhas de coleta, dutos de drenagem, poços de sucção (incluindo um Poço Geral de Recalque - PGR), *Holding Basins* – HB, equipamentos para tratamento preliminar de efluentes, tubovias para recalque e estações elevatórias são equipamentos que também constituem o sistema, que é operado remotamente.

As bacias de contenção pertencem a dois subsistemas e podem receber efluentes industriais de acordo com a presença eventual ou constante de hidrocarbonetos e contaminantes, denominados respectivamente como contaminados e oleosos (HODGSON; BENDIAK, 1987). O sistema de águas pluviais oriundas das zonas urbanas na refinaria não compõem o sistema de drenagem industrial.

Oito bacias compõem o sistema, sendo quatro para efluentes contaminados (CET 02, CET 03, CET 07 e CET 09) e outras quatro para efluentes oleosos (OET 02, OET 03, OET 07 e OET 09). Os principais elementos (quantidade de bombas de recalque na bacia – Nº de bombas; vazão média de entrada – Q_{in}; vazão de descarga – Q_{des}) e as características de cada bacia (capacidade volumétrica de cada bacia – Vol. Máx.), podem ser verificadas na Tabela 3.

Tabela 3 - Principais características das bacias de contenção em análise

Bacia	Nº de Bombas	Q_{in} [m³/h]	Q_{des} [m³/h.CMB]	Vol. Máx. [m³]
CET 02	02	10	35	640
CET 03	02	21	120	2 780
CET 07	02	68	120	2 903
CET 09	02	142	285	6 800
OET 02	02	167	400	10 760
OET 03	04	4	25	2 100
OET 07	02	24	90	192
OET 09	02	11	20	2 560

Fonte: O Autor, 2019.

Os efluentes contaminados são oriundos da parte externa da área de processos e circulação de produtos, já os efluentes oleosos são provenientes predominantemente de processos produtivos e de manutenção de equipamentos, como a lavagem de tanques. Ambos os subsistemas contam com tratamento preliminar para remoção de possíveis sólidos grosseiros presentes no efluente.

Quando as OETs estão totalmente cheias, os efluentes contidos nela podem transbordar efluente para as CETs ou para o corpo receptor. Além dessa comunicação entre bacias, existe também a situação de dependência entre bacias para o manejo adequado do sistema, como as CET 07 e CET 09. A hipótese de referência deste trabalho

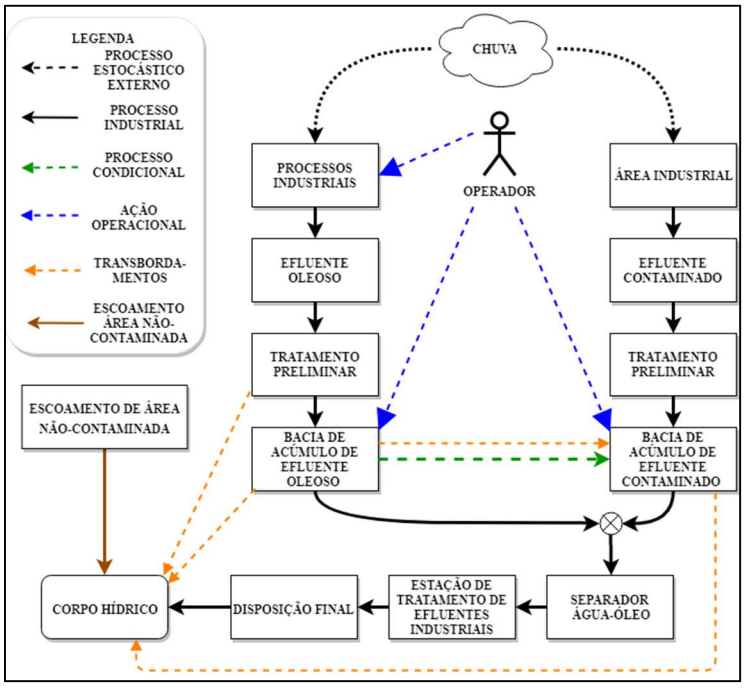
afirma ser possível diferenciar comportamentos operacionais típicos quanto ao uso dos componentes e itens do sistema pela série histórica de dados.

Todo o sistema de drenagem é monitorado e controlado em tempo real por operadores específicos da refinaria, como evidenciado na Figura 7. O monitoramento dos processos no sistema é realizado buscando garantir que as saídas do processo atendam aos requisitos necessários para operação visando também a segurança do processo (ROSEN; RÖTTORP; JEPPSSON, 2003).

O sistema de drenagem em questão é responsável por segregar, coletar e transportar os efluentes líquidos gerados dentro de toda área da refinaria, e seu principal objetivo é viabilizar a chegada do efluente nas unidades de tratamento de efluentes industriais.

Foram disponibilizados para esta pesquisa, dados de leituras de medição de nível destas bacias, dados do *status* dos conjuntos motor-bomba (CMB) e dados de precipitação da região para o período de 01/01/2011 à 14/09/2011. Os dados pluviométricos têm frequência diária, enquanto os dados de medição de nível e das bombas foram coletados com frequência de minuto.

Figura 7 - Sistema de Coleta, Contenção e Transporte de Efluentes Industriais em estudo.



Fonte: O Autor, 2019.

4.1 TRATAMENTO DE DADOS

Para o presente trabalho foi obtido junto à refinaria a série histórica do índice pluviométrico diário para a localidade da refinaria. O período em análise tem o perfil típico de chuvas referente à região frente a estimativa de normal climatológica analisadas (CLIMATEMPO, 2018; RAMOS; SANTOS; FORTES, 2009), satisfazendo a condição de que o sistema estava submetido a condições convencionais de operação. Do total das 257 observações, 154 dias com ocorrência de chuva foram observados, sendo que apenas para seis dias não havia informação quanto ao índice pluviométrico.

A modelagem aqui relatada foi desenvolvida utilizando o *software* R versão 3.5.1 (R CORE TEAM, 2018) e pacotes desenvolvidos por contribuidores.

O preenchimento de dados faltantes de precipitação foi realizado utilizando a técnica geoestatística de interpolação espacial a partir do método baseado ponderado pelo inverso da distância ao quadrado (COLLISCHONN; TASSI, 2008), Equação (3), utilizando o pacote *gstat* (PEBESMA, 2018).

Os dados utilizados na interpolação foram obtidos através de *web scraping* do Banco de Dados Meteorológicos para Ensino e Pesquisa (BDMEP) do Instituto Nacional de Meteorologia (INMET) por meio do pacote *inmetr* (TATSCH, 2018).

As estações pluviométricas utilizadas para obtenção de dados para modelagem quanto à precipitação em relação a região de localização da refinaria são referentes a quatro municípios próximos a região da refinaria. Esses municípios foram estrategicamente escolhidos em função da disponibilidade de dados e distribuição geográfica. O conjunto de dados de precipitação diária foi utilizado para toda a área da refinaria, desconsiderando a variação espacial de chuvas.

4.2 ANÁLISE DE COMPORTAMENTO E PADRÕES DE TRANSBORDAMENTO

4.2.1 Índice de similaridade para agrupamentos

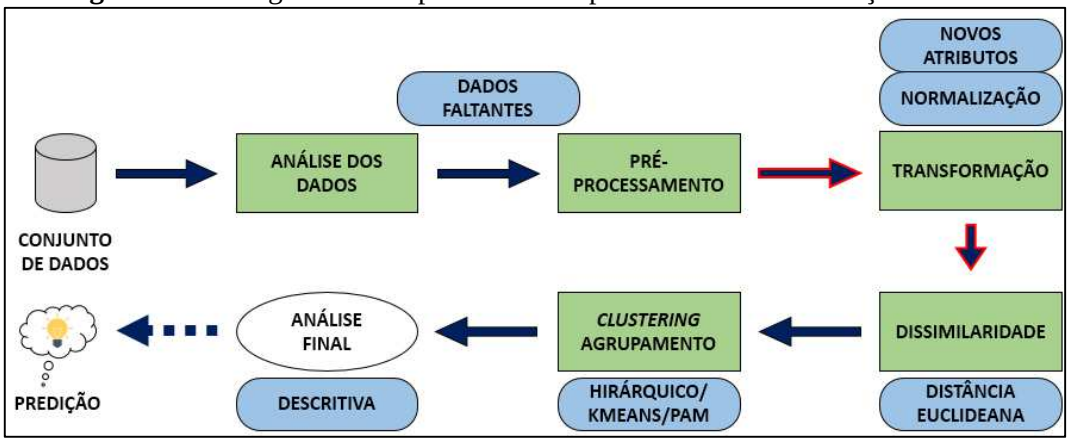
Objetivando investigar a relação entre as características da chuva e a ocorrência dos transbordamentos nas bacias de contenção, empregou-se a metodologia de similaridade entre grupos de transbordamentos de efluentes urbanos combinados com águas residuárias de drenagem e características hidrológicas (YU et al., 2013).

Dois tipos de agrupamentos foram conduzidos por dois conjuntos de atributos de acordo com o padrão de chuva e os comportamentos dos transbordamentos no sistema de drenagem industrial.

A informação dos dados foram transformadas para valores totais diários, onde cada dia de observação passou a ser um objeto. Sendo assim a informação contida representa o histórico de dados para cada dia no intervalo horário de 00:00hr à 23:59hr, totalizando 257 objetos.

A Figura 8 sintetiza as etapas que foram seguidas para realização do presente estudo, retratando desde o conjunto de dados iniciais coletados até os produtos finais de análise exploratória e construção do modelo preditivo.

Figura 8 – Fluxograma de etapas realizadas para análises de mineração de dados.



Fonte: O Autor, 2019.

O de pré-processamento de dados para inibir possíveis distorções na análise em função das diferenças da escala dos dados dos atributos, os valores dos atributos foram normalizados para escala *Z-score*.

Baseado nos atributos escolhidos e avaliados no estudo desenvolvido por Yu et al., (2013), no fenômeno chuva-vazão e nos dados e informações disponíveis, quatro atributos para agrupamento do comportamento relacionado com a ocorrência e frequência condicional (se-então) de chuvas foram escolhidos, sendo: Volume acumulado de chuva do dia anterior (Vc_{i-1} [mm/dia]); Volume acumulado de chuva de um dia (Vc_i [mm/dia]); Volume acumulado se ocorrido dois dias consecutivos chuva ($Vc_i + Vc_{i-1}$) [mm]; e Volume acumulado se ocorridos três dias consecutivos de chuva ($Vc_i + Vc_{i-1} + Vc_{i-2}$) [mm].

Para os atributos condicionais, caso não satisfeita a condição inicial dos volumes de chuva, foi considerado o volume 0 (zero) mm como resposta.

Quanto ao comportamento dos transbordamentos nas bacias, quatro atributos foram considerados para o conjunto, sendo que estes parâmetros representam o sistema como um todo, pois a ocorrência de transbordamento não acontece em todas as bacias e também existe variabilidade na quantidade de bacias com transbordamento.

Os atributos para o fenômeno transbordamento são análogos ao de sistema de drenagem urbana, sendo: Volume total de transbordamento por dia [m³/dia]; Número de bacias transbordando por evento; Tempo médio de transbordamento por dia [min/dia]; e Tempo médio diário de operação das bombas [min/dia].

O atributo relacionado à operação das bombas foi considerado em função da relevância e modelo de operação mecanizada das bacias, bem como o manejo dos efluentes na planta industrial. O volume total de transbordamento foi obtido pela soma do volume transbordado de cada bacia em cada dia, sendo estimado por meio de balanço de massa em cada bacia, representado na Equação 36 e Equação 37 como discutido por Gonzalez et al. (2013) e Oliveira-Esquerre et al. (2011).

$$\frac{dV(t)}{dt} = \begin{cases} Q_{in} - Q_{out} & ,se\ L < 99{,}8\% \\ Q_{out} = Q_{des} & \end{cases} \quad (36)$$

$$\begin{cases} 0 & ,se\ L > 99{,}8\% \\ Q_{out} = Q_{des} + Q_{tr} \therefore Q_{tr} = Q_{in} & \end{cases} \quad (37)$$

Onde *dV(t)/dt* é a taxa de variação do volume na bacia em relação ao tempo *t*; Q_{out} é a vazão de saída; *L* é o nível porcentual da bacia em um instante *t*; Q_{des} é a vazão da descarga referente as bombas; Q_{tr} é a vazão de transbordamento no instante *t*; Q_{in} é a vazão média de entrada na bacia. Deste modo, foi mensurado o volume transbordado conforme Equação 38:

$$Vol_{tr,bacia} = Q_{tr} \times t_{totaltr} \quad (38)$$

onde $Vol_{tr,bacia}$ é o volume transbordado por cada bacia em um dia e $t_{totaltr}$ é o tempo total em transbordamento de cada bacia em um dia. Quanto à medição de nível nas bacias e acionamento de bombas (quantidade de bombas ligadas/minuto) apesar de os instrumentos de medição apresentarem boa confiabilidade e baixa incerteza de medição,

alguns dados apresentaram erros ou ausências relativas ao sistema de coleta de dados ou a alguma falha no equipamento de medição. Contudo, essas falhas foram preenchidas utilizando a última informação íntegra da série temporal.

Os agrupamentos foram gerados baseados no cálculo de dissimilaridade da Distância Euclideana entre objetos, e com emprego de diferentes algoritmos, hierárquico com método de Ward, *K-means* e PAM (*Partitioning Around Medoids*) com o objetivo de escolher o melhor método de agrupamento e o número ótimo de gurpos. Para tal, foi realizada a verificação da qualidade e estabilidade dos agrupamentos utilizando o pacote *clValid* (BROCK et al., 2008).

A qualidade interna dos agrupamentos foi verificada através de três medidas: Índice Dunn, Conectividade e Silhueta. Para a análise da estabilidade dos agrupamentos empregou-se as técnicas de proporção média de observações não classificadas no mesmo grupo (*average porportion of nonoverlap* - APN), distância média entre objetos em um mesmo grupo (*average distance* - AD), distância média entre os centroides quando as observações estão no mesmo grupo (*average distance between means* - ADM) e figura de mérito (*figure of merit* - FOM).

Foi analisado também o Índice de Similaridade entre os dois grupos categorizados de precipitação e transbordamento proposto por Yu et al. (2013).

4.2.2 *Change-point*

No presente trabalho foi também empregada a análise de *change-point* considerando os parâmetros média e variância para detecção de mudanças nas séries temporais referente à precipitação e ao nível das bacias, sendo realizado o teste de hipótese conforme a Equação (21) e Equação (22) (COSTA; GONÇALVES; TEIXEIRA, 2016). A modelagem foi realizada utilizando o pacote *changepoint* (KILLICK; ECKLEY, 2014).

Utilizou-se o algoritmo de Segmentação Binária com critério de penalidade bayesiano SIC (Critério de informação de Schwarz), para caracterizar a probabilidade de ajuste dos modelos de segmentação testados para as séries temporais, devido à relevância da não necessidade de atendimento aos pressupostos de estacionariedade, normalidade e ausência de autocorrelação (COSTA; GONÇALVES; TEIXEIRA, 2016).

Como a informação da precipitação pluviométrica está em escala diária, convencionou-se transformar cada série de nível das bacias [% nível/min.] em mediana

do percentual de nível diário, de modo a reduzir o número de observações de cada série temporal tendo em vista a manutenção quanto ao comportamento de cada série ao longo do tempo.

Deste modo, diferentes números de segmentos foram encontrados para cada série univariada. A determinação do número de change-points para cada série foi realizada utilizando máxima verossimilhança, com o conjunto sequencial mínimo de duas observações que foi considerado como comprimento mínimo do segmento.

4.3 MODELAGEM PREDITIVA PARA TRANSBORDAMENTO

Considerando as informações previamente analisadas na abordagem não supervisionada, a análise de confiabilidade realizada sobre o sistema (SANTANA; PESSOA; OLIVEIRA-ESQUERRE, 2017), e a modelagem de simulação hidrológica (GONZALEZ et al., 2013), foi realizada a modelagem preditiva para transbordamento da principal bacia de contenção de efluente contaminado, a CET 09.

A bacia de contenção CET 09, conforme Tabela 03, recebe em média uma vazão de entrada (Q_{in}) de 142 m³/h, tem capacidade útil de armazenamento (VolMáx.) de 6 800 m³ de efluente, contendo também dois conjuntos motor-bomba (CMB) que proporcionam uma vazão de recalque (Q_{dis}) de 285 m³/h. CMB.

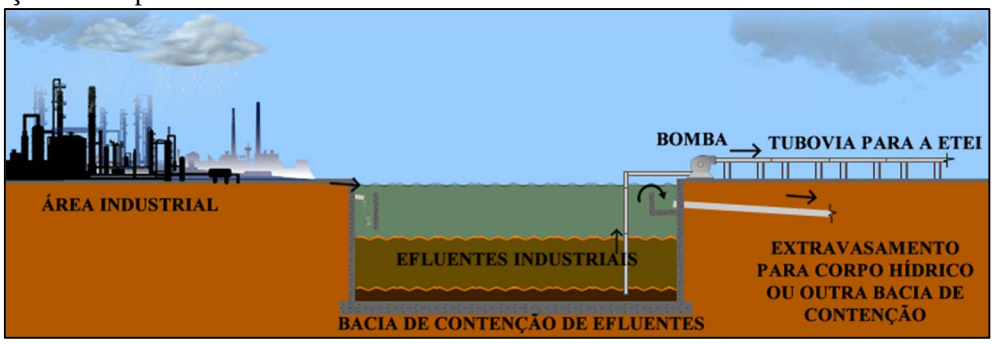
Os dados brutos utilizados para a modelagem aqui retratada foram tratados de maneira análoga à abordagem não supervisionada, contudo as características retratadas para a predição refletem especificamente para a bacia CET 09. Na Figura 9 é possível visualizar um croqui a respeito da CET 09 e variáveis de interesse, como a entrada do efluente e possíveis saídas.

A modelagem preditiva aqui realizada é a de classificação para o dia posterior, pautada na predição da ocorrência ou não de transbordamento de efluente, com abordagem considerando cada dia observado como um objeto. Ou seja, comparando com a abordagem não supervisionada os dados para predição sofreram defasagem de um dia, exceto os rótulos, a fim de proporcionar horizonte de predição.

Foram construídos vinte e quatro modelos de aprendizado de máquina, sendo doze modelos utilizando *Random Forest* e outros doze com KNN. Os doze modelos para cada algoritmo foram divididos em três cenários, onde cada cenário possuí um conjunto

específico de variáveis independentes. Também foram aplicadas técnicas de reamostragem para compensação de desbalanceamento de informação de classes.

Figura 9 – Croqui, sem escala, representando a bacia CET 09 no sistema de coleta, contenção e transporte de efluentes industriais em estudo.



Fonte: O Autor, 2019.

Para minimizar possíveis viés, para cada algoritmo e cenário foram construídos modelos preditivos considerando os dados reais e os dados com reamostragem, sendo empregado as técnicas de reamostragem de *under-sampling*, *over-sampling* e ROSE.

O conjunto de variáveis do Cenário 01 é baseado no fenômeno chuva-vazão, e para construção desse modelo foi utilizada apenas a informação de precipitação pluviométrica diária (Vc_{i-1}) acumulada [mm/dia] no dia anterior e variáveis relacionadas ao nível da CET 09, a saber: mediana diária do nível, máximo nível diário, desvio-padrão do nível diário e a diferença do range do nível diário.

Para o Cenário 2, foram utilizadas as mesmas variáveis do Cenário 1 com o incremento de variáveis relacionadas à ocorrência de chuvas. As variáveis inseridas são: volume de precipitação pluviométrica com defasagem de dois dias (Vc_{i-2}) [mm], e variáveis condicionais (se, então), como volume acumulado se ocorrido dois dias consecutivos de chuva ($Vc_{i-1} + Vc_{i-2}$), volume acumulado se ocorrido três dias consecutivos de chuva ($Vc_{i-1} + Vci\text{-}2 + Vc_{i-3}$) e informação do processo operacional.

Para as variáveis condicionais, em caso de ausência de dias consecutivos de chuva, a informação é zero. A informação do processo operacional é a variável tempo diário de descarga, ou seja, o tempo diário que o total de bombas ficou ligado durante o dia. Esta informação é necessária devido a existência dos dois CMB na CET 09, pois a operação ótima e a prática eficaz podem reduzir os riscos de inundação e transbordamentos (YAZDI; CHOI; KIM, 2016).

Logo, o tempo diário é o somatório de tempo de trabalho de cada bomba. A informação da precipitação com defasagem de dois dias é importante devido à baixa resolução dos dados de precipitação, que estão apenas a nível diário. A chuva acumulada a nível diário dificulta a compreensão quanto a ocorrência e implicação no fenômeno chuva-vazão, uma vez que não se conhece o tempo de concentração da bacia hidráulica CET 09.

O Cenário 3 são as variáveis do Cenário 2 com a adição de variáveis relacionadas ao fenômeno de transbordamento. As variáveis adicionais ao Cenário 3 são: Tempo diário da CET 09 em transbordamento [min.], volume diário de transbordamento [m³] e máximo nível percentual diário (*i-1*) no intervalo último intervalo horário de cada objeto, compreendendo de 23:00hr à 23:59hr.

A variável referente ao último intervalo horário do objeto tem como objetivo estruturar informação de modo a tentar captar o comportamento do nível da bacia de contenção no período de transição entre os dias, objetos. O período de 23:00hr à 23:59hr é utilizado porque é o último intervalo horário de 00:00hr à 23:59hr do período de cada objeto.

Das variáveis incrementadas, o tempo e o volume de transbordamento são variáveis que foram amplamente pesquisadas por Yu et al. (2013, 2018), estas variáveis tendem a proporcionar a informação de resposta para variáveis como intensidade e volume de chuvas.

A variável do máximo nível no último intervalo horário do dia foi escolhida em função da possibilidade de trabalho com a observação da frequência de transbordamento em intervalos de hora e da característica de predição diária.

O volume diário de transbordamento foi mensurado realizando balanço de massa, Equação (37), com um abordagem mais conservadora, considerando que em situações onde o nível da bacia é maior que 95% e o conjunto motor-bomba está desligado, ocorre o transbordamento. Essa premissa permite proporcionar maior segurança para a operação. Informações sobre a qualidade do efluente e possíveis implicações a jusante do sistema não foram consideradas para a modelagem.

Para a construção dos modelos os dados foram pré-processados utilizando a padronização *Z-score* com centralização, onde cada modelo foi dividido em 75% dos

dados para treinamento e 25% para teste. Para validação estatística foi realizado a validação cruzada *k-fold*, sendo $k = 10$.

4.3.1 Avaliação dos modelos de predição

A predição para classificação utilizando KNN e *Random Forest* fornece resultados probabilísticos (votos) para as classes. E, no presente trabalho convencionou-se utilizar o *threshold* (limiar) padrão, onde para discriminação de dois grupos considera-se como regra de classificação a probabilidade de ocorrência de transbordamento maior que 50%, caso contrário classifica-se como operação normal.

Para avaliar o desempenho dos modelos preditivos construídos foram utilizados os scores da matriz de confusão de cada modelo, conforme exemplificado na Tabela 2.

Delineado o evento de interesse, que é o fenômeno de transbordamento da CET 09, a predição correta deste é um verdadeiro positivo (VP), FP é um falso positivo (quando o modelo incorretamente prevê transbordamento), VN é um verdadeiro negativo (corretamente predito a operação normal operação), e FN é um falso negativo (quando há a predição incorreta de estado normal de operação).

Organizando as informações por categorias preditas e observadas na matriz de confusão, foi empregado as métricas de Acurácia, Sensibilidade, Especificidade, *Kappa* (COHEN, 1960), *F1 Score* e ORSS *(Odds Ratio Skill Score* – índice de taxa de probabilidade) para avaliação da escolha dos modelos.

O algoritmo KNN foi implementado utilizando Distância Euclideana para o cálculo de similaridade entre objetos, e o número de vizinhos mais próximos (K) compreende o intervalo de $k = \{1, 2, ..., 24, 25\} \in N$, que é um intervalo que engloba o possível número de vizinhos ótimos conhecidos empiricamente (HASSANAT; ABBADI; ALTARAWNEH, 2014; ZHANG et al., 2017).

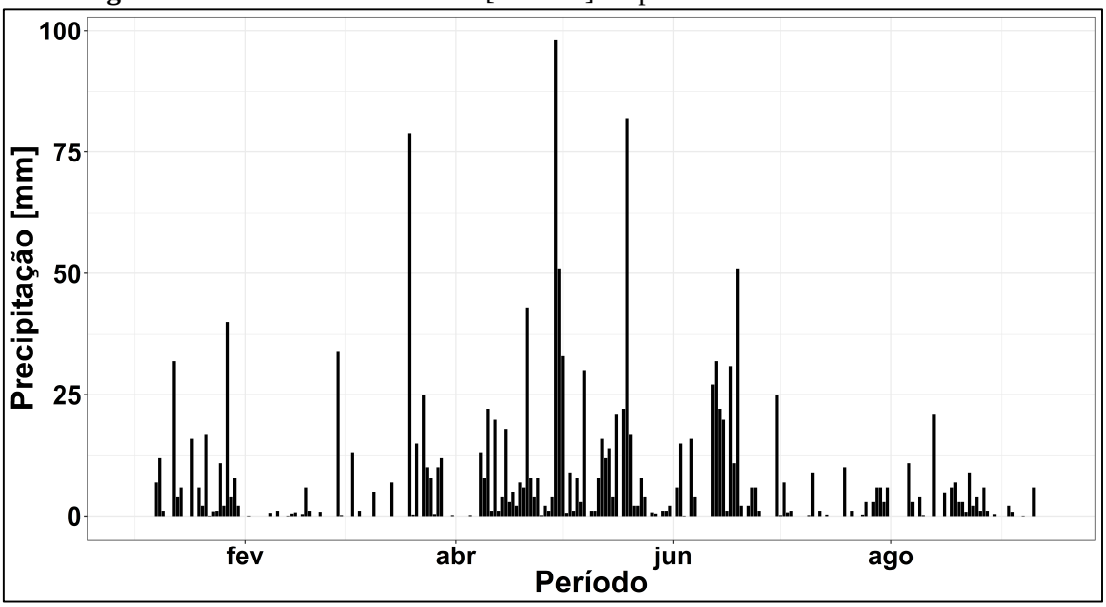
Para o algoritmo *Random Forest* foi considerado o número de 1500 árvores aleatórias. Já o número de preditores (*mtry*), disponível para divisão em cada nó das árvore, foi concebido pelo pacote *caret* (KUHN, 2018).

5 RESULTADOS

5.1 ANÁLISE EXPLORATÓRIA

Dos 154 objetos com ocorrência de chuvas, a precipitação máxima diária ocorrida foi de 98,0mm e a precipitação mínima diária foi de 0,1mm, como pode ser visualizado na Figura 10. Aproximadamente 28,6% dos objetos chuvosos implicaram em ocorrência de transbordamento, no qual o menor volume de efluente extravasado diário, neste cenário, foi de 0,67m³, com precipitação diária de 2,0mm.

Figura 10 – Índice Pluviométrico [mm/dia] no período em análise.



Fonte: O Autor, 2019.

Os maiores volumes de efluente extravasado diário (2 385,50m³ e 2 380,80m³) ocorreram em objetos, dias, chuvosos e relativamente com baixa precipitação diária, respectivamente 2,0mm e 9,0mm, e baixa precipitação acumulada diária considerando a informação de objetos anteriores, respectivamente 11,8mm e 12,0mm para três objetos consecutivos de chuva.

Nesses dois objetos o tempo de trabalho das bombas das bacias estava abaixo da média (aproximadamente 524 minutos/dia), sendo respectivamente 373 minutos/dia e 187 minutos/dia.

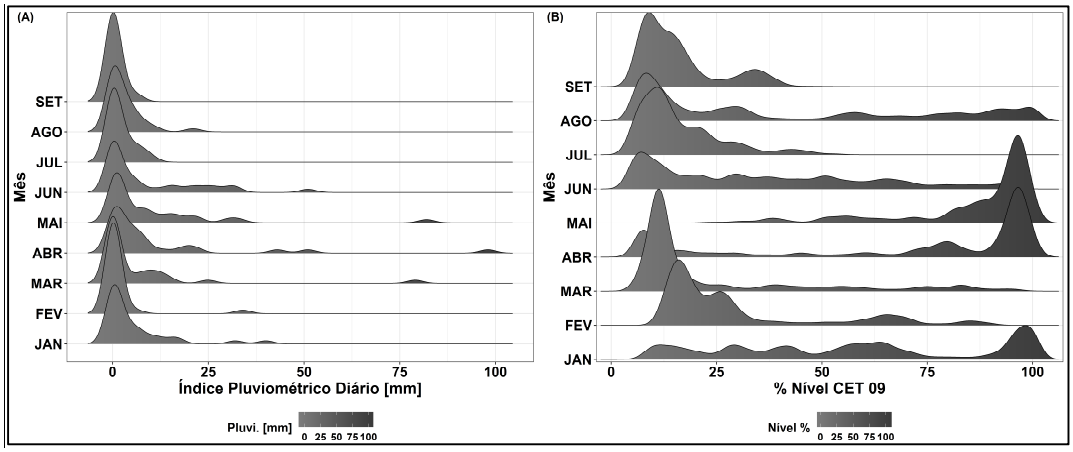
Em tempo seco, em 15 dos 257 objetos ocorreu transbordamento no sistema, sendo o volume máximo diário de 185,17m³ e o volume mínimo diário 0,5m³, tendo volume médio diário de transbordamento 30,34m³ e a mediana diária de 10m³. Para este

maior volume de transbordamento, o tempo médio de trabalho das bombas foi de apenas 277 minutos.

Os meses com maior índice pluviométrico acumulado são de abril à junho, e neste período contém 66% do registros das ocorrências de bacias em transbordamento, inclusive ocorrendo no dia da máxima precipitação registrada neste estudo.

Verificou-se também que nos meses centrais do ano há uma maior distribuição de chuvas com tendência a índice pluviométrico diário de 25mm. O período que compreende os meses de Abril à Junho abrange os meses onde ocorrem as maiores índices pluviométricos, como também pode ser interpretado na Figura 11.

Figura 11 – Distribuições de densidade. (A) Chuva mensal [mm]. (B) Percentual % de nível mensal na CET 09.



Fonte: O Autor, 2019.

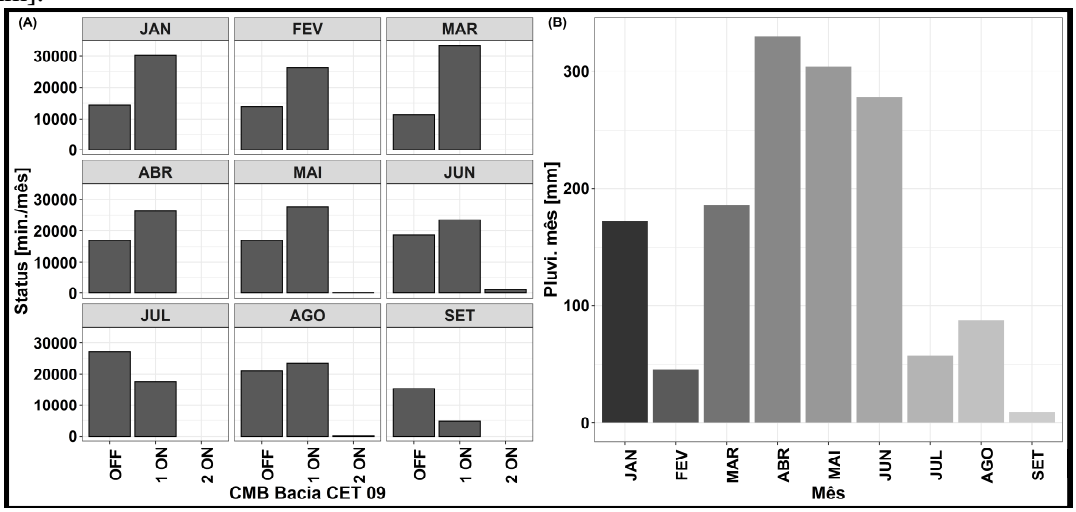
O nível da CET 09 tem um outro perfil de distribuição, o nível da bacia pode ser estratificada em níveis baixos e níveis altos, indicando que níveis centrais são apenas transitórios, ou seja, não tendo um período de estacionariedade para estes.

Na Figura 11 constatou-se também que a mudança da assimetria da distribuição do nível da CET 09 ocorre de maneira mais acentuada nos meses de Abril e Maio. Meses como Janeiro e Agosto apresentam uma distribuição com maior curtose.

Considerado que o tempo de recorrência *TR* foi de 20 anos para vazão de projeto do sistemas, o sistema tem uma probabilidade projetada de falha em 5%, que representa 18,25 falhas ao longo de um ano (365 dias). Entretanto, nos 257 objetos observados, mensurou-se 44 (17%) falhas em tempo chuvoso, e 15 falhas em tempo seco (5,8%), totalizando 59 falhas.

Quanto à operação do sistema de drenagem, apenas os meses de Julho e Setembro os conjuntos motor-bomba tem um tempo maior em situação desligada que o tempo em trabalho, Figura 12. Tal fato reflete a redução dos dias de chuvas nesses meses durante esse período central do ano.

Figura 12 – (A) Frequências acumuladas observadas de operação das bombas na CET 09. (B) Frequências acumuladas observadas de volume acumulado precipitação pluviométrica [mm].



Fonte: O Autor, 2019.

Verificou-se também, através do balanço de massa, que a ocorrência de transbordamentos tem a maior frequência predominantemente no período que compreende as maiores índices pluviométricos [mm/dia].

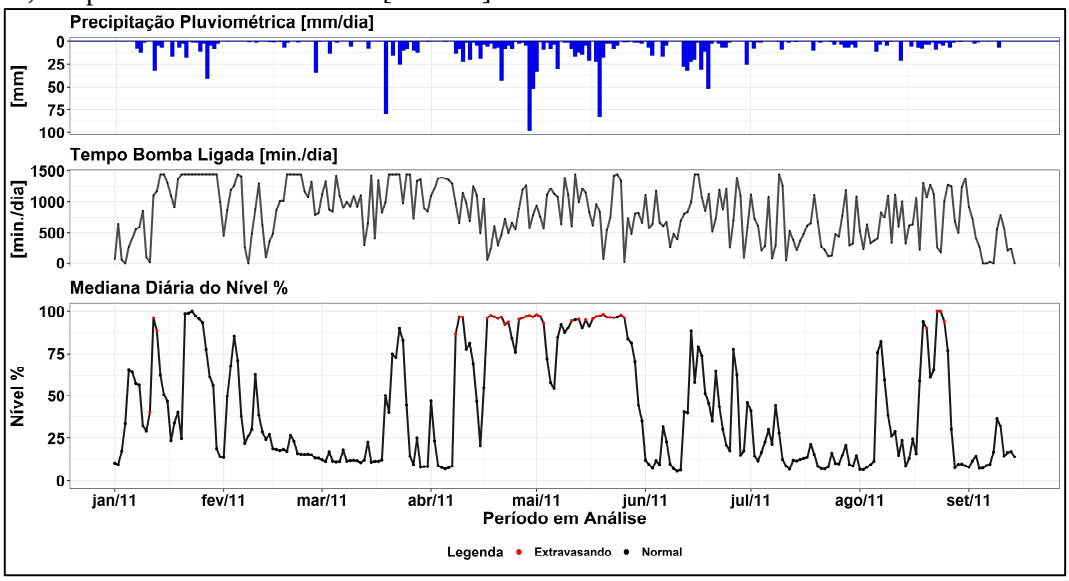
Nota-se que há uma redução do tempo de trabalho das bombas da CET 09 nos períodos em questão, Figura 13. Essa não conformidade operacional reflete possivelmente a importância de ajustar a operação CMB de forma a reduzir os transbordamentos, como sinalizado por Yazdi; Choi; Kim, (2016).

Verifica-se também que a representação da série histórica do nível transformado para mediana diária é uma boa representação do comportamento do fenômeno, apesar existir algumas regiões com picos mais intensos. Isto acontece devido a assimetria da distribuição de probabilidade.

Com a análise exploratória também foi possível avaliar o desempenho da CET 09 conforme metodologia adaptada a partir de Montserrat et al. (2015) construindo os gráficos do índice pluviométrico [mm] *versus* informações sobre transbordamentos, Figura 14. Constatou-se que ocorreu transbordamento em dias com índice pluviométrico

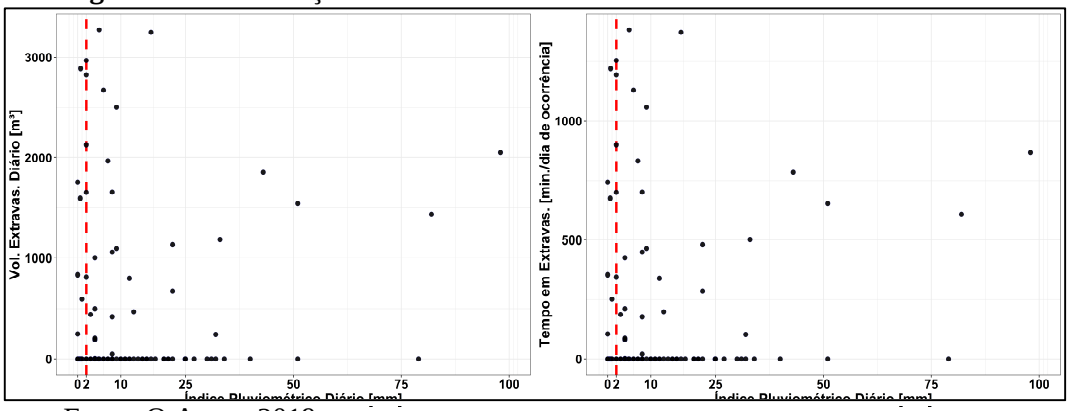
menor que 2 mm, caracterizando que a CET 09 não foi operada ao ponto de mitigar satisfatoriamente o incremento de vazão em episódios de chuva.

Figura 13 – Séries temporais de índice pluviométrico [mm/dia], mediana do % do nível, tempo de trabalho dos CMB [min/dia].



Fonte: O Autor, 2019.

Figura 14 – Verificação do limiar de transbordamento na CET 09.



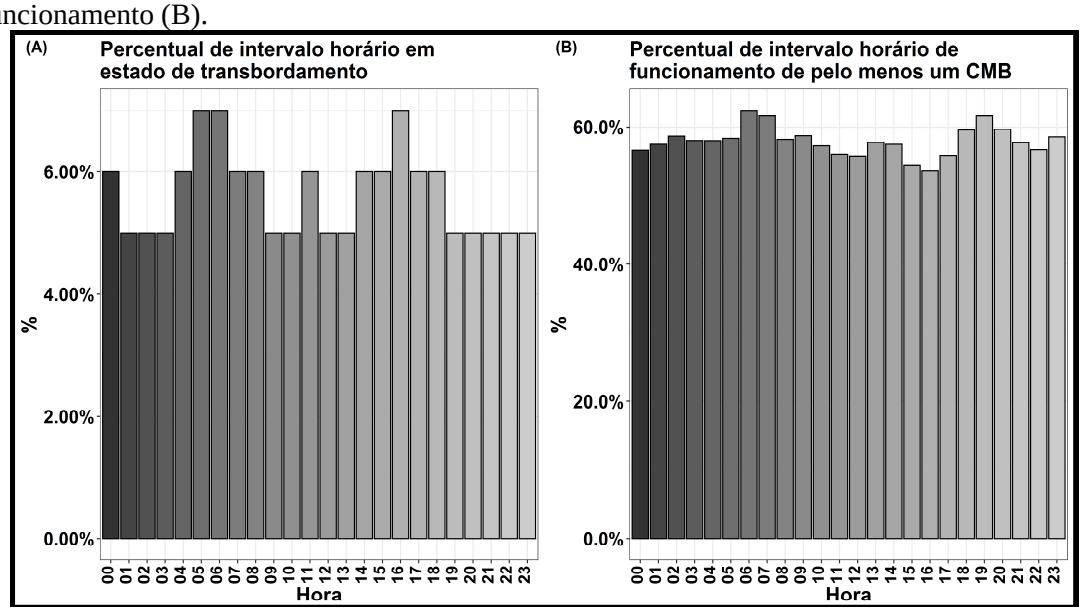
Fonte: O Autor, 2019.

A ausência de padrão de transbordamento, como pode ser visualizado na Figura 14, frente a metodologia de Montserrat et al. (2015), reflete a não linearidade e complexidade do sistema.

Tendo em vista a possibilidade de derivação de atributos de modo a facilitar a análise dos modelos foi verificado a percentual de transbordamentos durante o intervalo horário ao longo de uma dia, ou seja, $I \in \{0 \dashv 1; 1 \dashv 2; \dots; 22 \dashv 23; 23 \dashv 24\ hora\}$, e pode-se constatar através da Figura 15 que o horário limítrofe diário, 23⊣24 hora, possuí elevada frequência de transbordamentos, indicando que este período transitório pode

carregar uma importante informação quanto ocorrência de transbordamento de um dia para o outro.

Figura 15 – Percentual de transbordamentos por hora (A) e de pelo menos um CMB em funcionamento (B).



Fonte: O Autor, 2019.

Outros intervalos horários apresentam frequência de extravasamento maior, como 00 ⊣01 hora, ocorreram 956 minutos de transbordamentos, que retorna ~6% dos eventos, do total dos 15 420 minutos para cada intervalo horário.

Intervalos como 04 ⊣05 hora e 16 ⊣17 hora, por não estarem presentes na região marginal do dia, para a modelagem aqui realizada, a informação não apresenta tanta relevância para análise a nível diário, mesmo com percentuais mais elevados de transbordamentos.

A exploração dos dados auxiliaram na definição de variáveis da preditoras para a modelagem e criação dos cenários, que foram previamente apresentadas na metodologia, bem como na melhor compreensão acerca do fenômeno em análise na refinaria.

Podemos constatar que variáveis como valor máximo, mediana e média referentes ao nível diário das 23 horas às 24 horas, são possíveis covariáveis do modelo preditor diário de transbordamento, tendo em vista a possibilidade de interpretação de que transbordamentos estão ocorrendo de um dia para o outro.

5.2 ANÁLISE DE COMPORTAMENTO DE OPERAÇÃO DAS BACIAS

Na análise de agrupamento, para o agrupamento dos dados de chuva os resultados da análise de validação interna indicam que o agrupamento hierárquico com método de Ward foi o melhor nos três critérios, como pode ser visto na Tabela 4. Para a métrica de conectividade e silhueta, dois grupos foram indicados, já para o índice Dunn, cinco grupos foram indicados. Nos Apêndices é possível visualizar todos os resultados de validação interna dos agrupamentos através de figuras.

Tabela 4 - Métricas para validação interna de agrupamentos

Conjunto de Dados	Métrica	Ótimo *Score*	Melhor Algoritmo	Grupos
Chuva	Cnc	8,7698	Hierárquico	2
Chuva	Sil	0,8298	Hierárquico	2
Chuva	Dunn	0,5463	Hierárquico	5
Transbordamento	Cnc	0,0000	PAM	2
Transbordamento	Sil	0,8259	Hierárquico	2
Transbordamento	Dunn	1,0571	Hierárquico	2

Legenda: Cnc – Conectividade; Sil – Silhueta, Dunn – Índice Dunn.
Fonte: O Autor, 2019.

Em relação à estabilidade dos agrupamentos, as estatísticas apresentadas na Tabela 5. Para os dados relacionados a chuva, não conotam uniformidade, porém, dos quatro resultados, o método hierárquico apresenta melhor resultado para proporção média de observações não classificadas no mesmo grupo (APN) e distância média entre os centroides quando as observações estão no mesmo grupo (ADM), embora o número de grupos seja de cinco. Nos Apêndices é possível visualizar todos os resultados de estabilidade de agrupamento através de figuras.

Tabela 5 - Métricas para verificação de estabilidade dos grupos.

Conjunto de Dados	Métrica	Ótimo *Score*	Melhor Algoritmo	Grupos
Chuva	APN	0,0055	Hierárquico	2
Chuva	AD	0,8120	PAM	2
Chuva	ADM	0,0753	Hierárquico	5
Chuva	FOM	0,6124	*k-means*	5
Transbordamento	APN	0,0222	Hierárquico	2
Transbordamento	AD	0,8952	PAM	2
Transbordamento	ADM	0,1211	Hierárquico	2
Transbordamento	FOM	0,7386	*k-means*	2

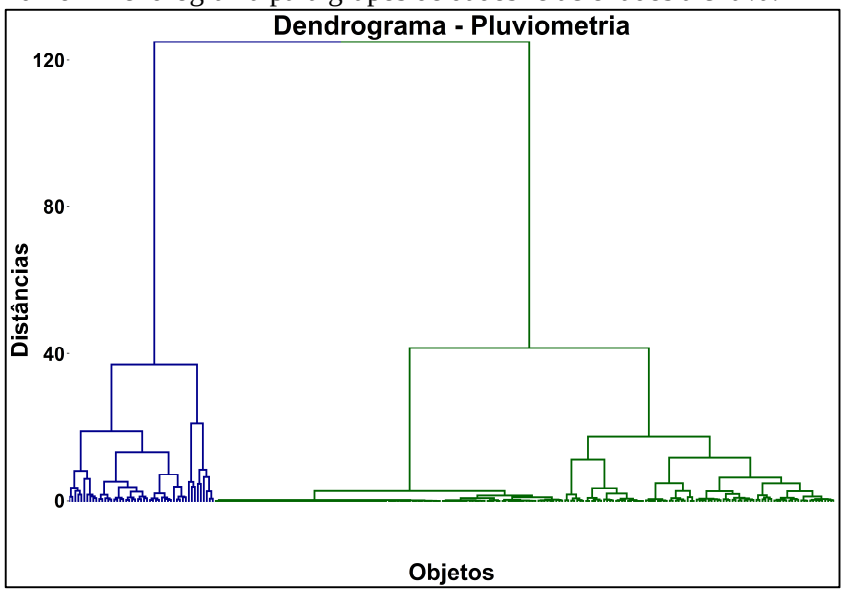
Fonte: O Autor, 2019.

Quanto ao agrupamento dos dados relacionados ao fenômeno de transbordamento, verificou-se através da validação interna que o método de agrupamento hierárquico com dois grupos é o de maior frequência, conforme pode ser visto na Tabela 4.

Verificou-se com a análise das estatísticas para estabilidade uma situação semelhante ao agrupamento dos dados de chuva, indicando o método hierárquico de Ward como o melhor método para agrupamento.

Isto foi elucidado pelos valores de APN e ADM, estabelecendo melhor estabilidade com agrupamento em dois grupos, o que corrobora com a análise da validação interna, também podendo ser certificado a partir da Figura 16.

Figura 16 – Dendrograma para grupos de dados relacionados a chuva.



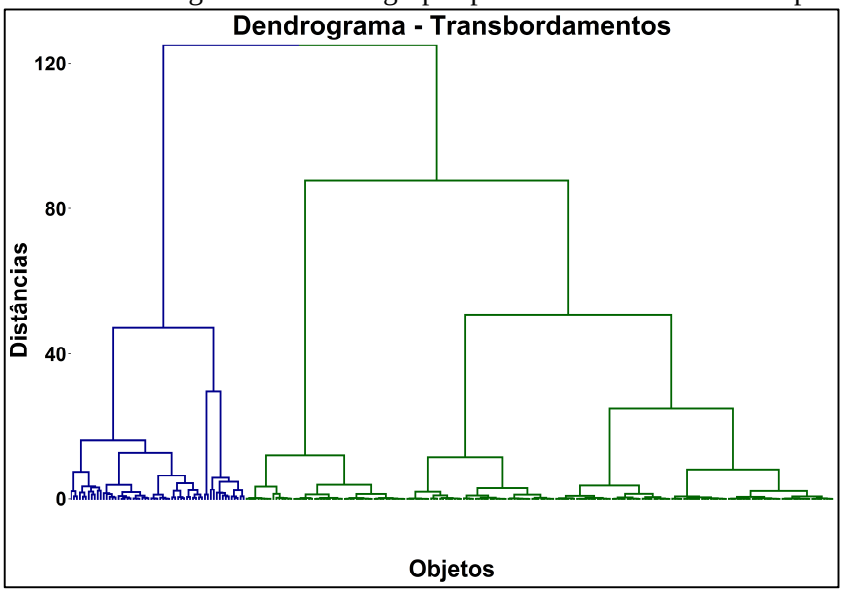
Fonte: O Autor, 2019.

Logo, a partir dos resultados obtidos na validação e verificação dos agrupamentos indica-se que a análise deve ser feita com dois grupos para os dados de chuva, e outros dois grupos para os dados de transbordamento, ambos utilizando o algoritmo de agrupamento hierárquico com método de Ward. Ambos os agrupamentos hierárquicos podem ser visualizados a partir dos respectivos dendrogramas na Figura 16 e Figura 17.

A partir destes resultados quanto ao emprego do algoritmo de agrupamento hierárquico para grupos para a série histórica de precipitação (CC) e os grupos para o fenômeno de transbordamento (CE), reafirmam que o método de Ward é o melhor método de agrupamento para análise de fenômenos físicos (YU et al., 2013).

Contudo, ao analisar o dendrograma na Figura 17, através de heurística, é latente a possibilidade de também realizar o agrupamento dos dados referente a transbordamento com até três grupos, tendo em vista a ampla distância para tal subdivisão.

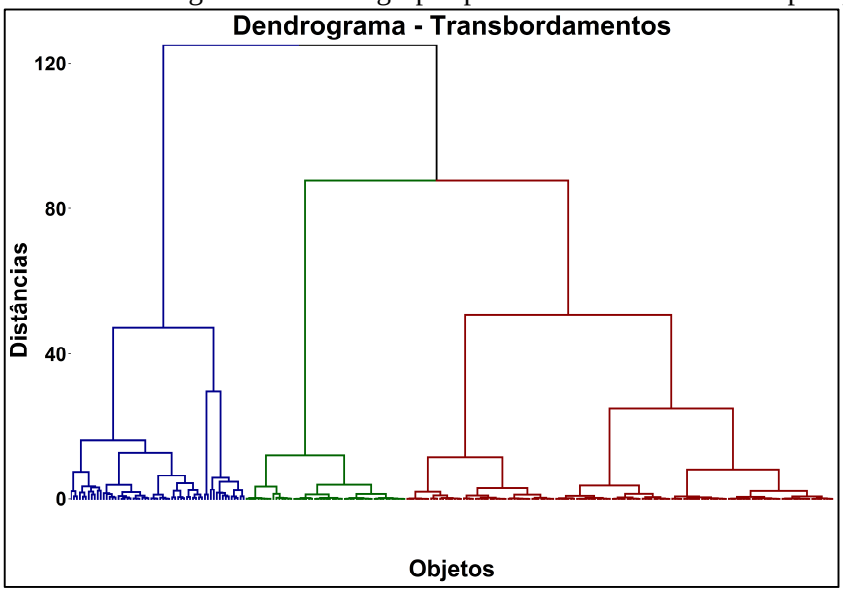
Figura 17 – Dendrograma com dois grupos para dados relacionados a operação.



Fonte: O Autor, 2019.

Deste modo, foram corridas duas análises acerca do Índice de Similaridade (IS). Incialmente, dois grupos foram considerados quanto aos dados de chuva e outros dois grupos para os dados de transbordamento, e em outra avaliação considerou-se a heurística, percepção do pesquisador para divisão de grupos, e adotou-se três grupos para os dados de transbordamento, como pode ser visualizado na Figura 18.

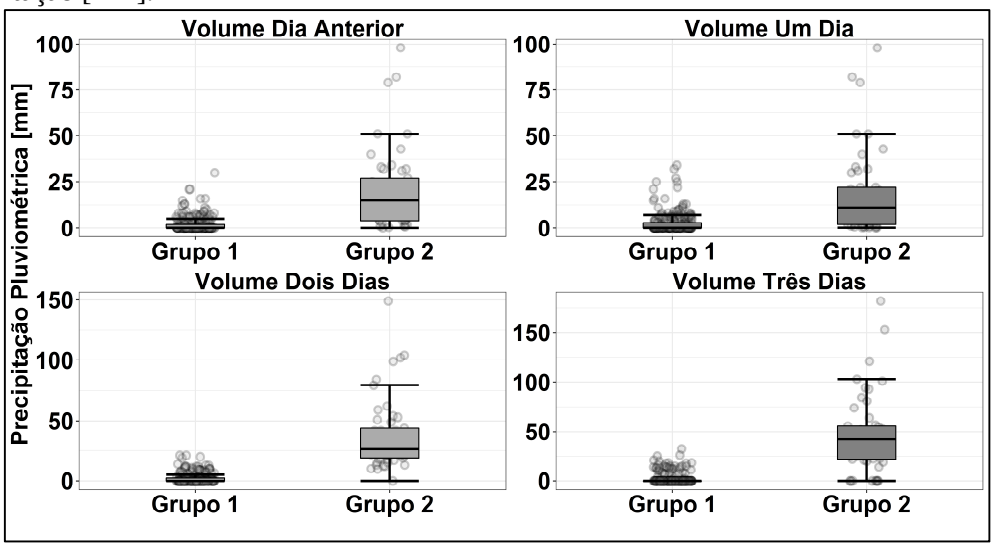
Figura 18 – Dendrograma com três grupos para dados relacionados a operação.



Fonte: O Autor, 2019.

No agrupamento dos dados referente a ocorrência de chuvas, 208 objetos estão contidos no grupo um (CC 1) e 49 objetos estão contidos no grupo dois (CC 2), sendo representado na Figura 19 em função dos atributos avaliados.

Figura 19 – Grupos relacionados à ocorrência da precipitação e atributos em função da precipitação [mm].



Fonte: O Autor, 2019.

No grupo 1 dos dados relacionados a precipitação (CC 1) em aproximadamente 50% dos objetos há ausência de chuva no dia um ou no dia anterior. Considerando 2 mm como um limiar do índice pluviométrico baixo (MONTSERRAT et al., 2015), em 72% dos objetos há ocorrência de chuvas baixas, podendo este grupo ser caracterizado como característico de baixas ou ausente precipitação pluviométrica.

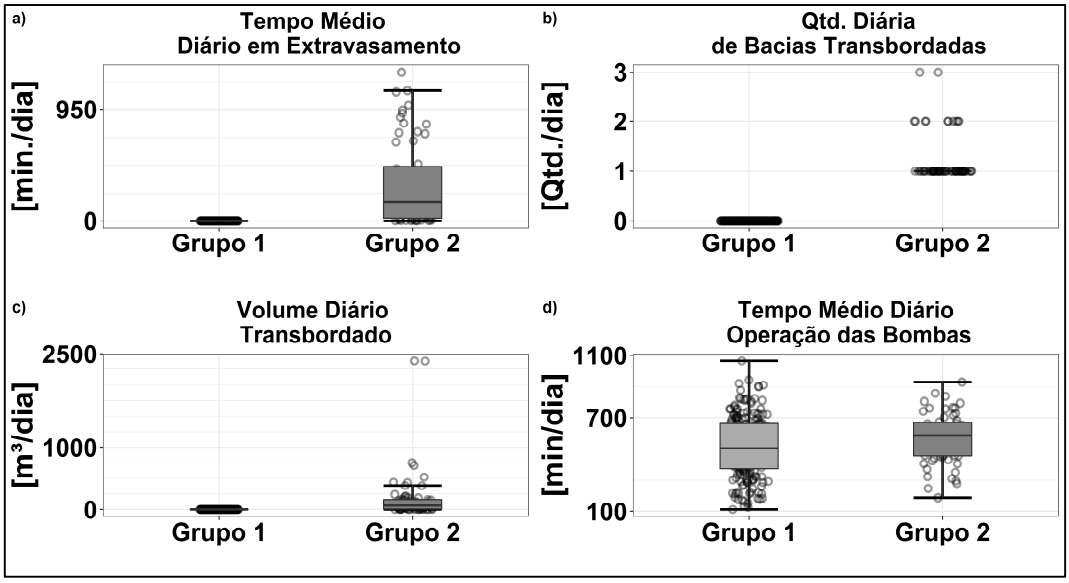
Apenas 20% dos objetos do CC 1 apresentam ocorrência de três dias consecutivos de chuva, sendo o maior volume acumulado de 32,2 mm. E em treze objetos ocorreram o volume de chuva maior ou igual a 10 mm para um dia sem ocorrência em dias sequenciais, sendo o maior volume de 32 mm. Tal comportamento é diferente no grupo 2 dos dados relacionados a precipitação (CC 2), onde há apenas um objeto com situação similar, entretanto, com índice pluviométrico de 79 mm.

O CC 2 conta com 84% dos objetos com ocorrência de chuva em até três dias consecutivos, em 48 dos 49 objetos há ocorrência de chuvas consecutivas em até dois dias, e em todos objetos há ocorrência de chuva diária, caracterizando este grupo como os objetos com dias chuvosos, sendo 53% da chuva diária maior ou igual que 10 mm.

Com relação à análise operacional e ocorrência de transbordamentos, dois distintos agrupamentos foram estabelecidos, com dois grupos baseados nos critérios de

validação interna e estabilidade (CED) e com três grupos baseado na heurística (CETH), sendo apresentados na Figura 20 e Figura 21.

Figura 20 – Grupos CED relacionados aos transbordamentos no sistema de drenagem de efluentes industriais e atributos em análise.



Fonte: O Autor, 2019.

Do agrupamento dois grupos baseados nos critérios de validação interna e estabilidade (CED), percebe-se claramente que a ocorrência de transbordamentos está contida apenas no grupo CED 2, sendo notório que em média apenas uma bacia transborda a cada evento.

A bacia CET 02 pode ser considerada a bacia mais vulnerável do sistema, pois está presente em 83% dos 59 objetos com ocorrência de transbordamento, tal característica ocorre em função do pequeno volume de acumulação frente as outras CET, como mostrado na Tabela 3. As outras bacias que apresentaram transbordamento são CET 09, CET 07 e OET 07, elencadas em ordem crescente de quantidade de transbordamento.

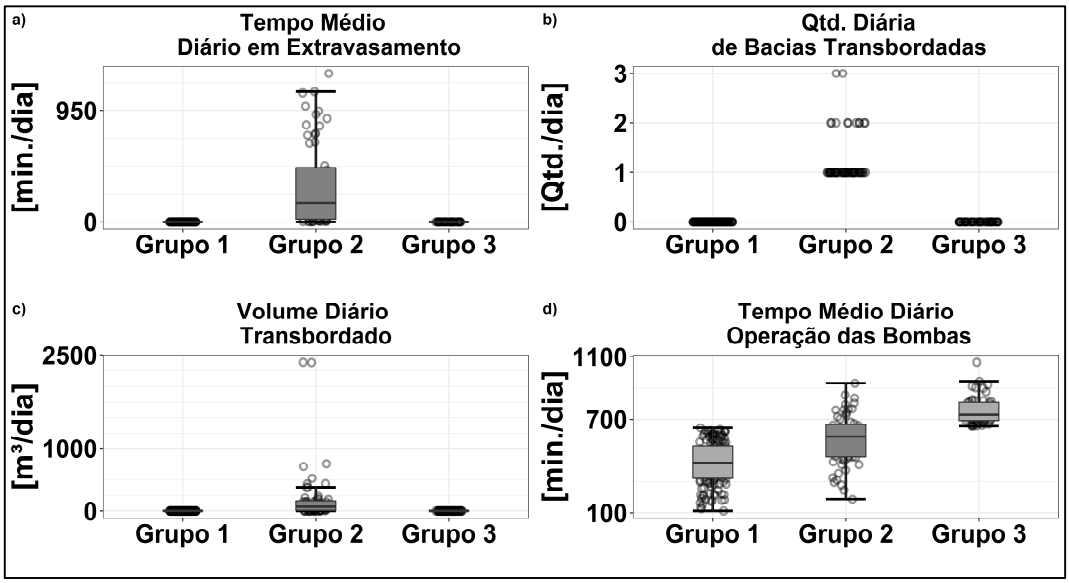
Dos 59 objetos que registraram transbordamento, 73% aconteceram na circunstância de ocorrência de chuva no dia anterior, 57% para dois dias consecutivos de chuva, e 47% aconteceram no cenário com três dias de chuva consecutiva acumulada.

Dos atributos categorizados para a operação das bacias, no grupo CED 2 observa-se maior variabilidade e maior média quanto à operação das bombas no sistemas, evidenciando um maior aporte do sistema para solucionar situações com elevado nível nas bacias de contenção.

Do grupo CED 2, apenas 12 objetos contém a informação de duas ou três bacias transbordando, e somente um dia de tempo seco (21/02) registra a informação de duas bacias transbordando, pertencendo ao grupo CED 1 e grupo CC 2.

Quando realizado o agrupamento em três grupos baseado na heurística, CETH, o maior grupo baseado nos critérios de validação interna e estabilidade, o CED 1, que contém 198 objetos, é decomposto em dois grupos, sendo o CETH 1 com 144 objetos e o CETH 3 com 54 objetos. O CETH 2 é idêntico ao CED 2, contendo os 59 objetos restantes.

Figura 21 – Grupos CETH relacionados aos transbordamentos no sistema de drenagem de efluentes industriais e atributos em análise.



Fonte: O Autor, 2019.

A conjuntura é complexa, mas é possível constatar, em conjunto com a análise descritiva, que se o sistema de contenção for operado e manejado de maneira com estratégia mais contínua, maior tempo de trabalho, reduz-se a possibilidade de acontecimentos de transbordamentos. Sendo este resultado similar ao elucidado por Kusch; Haag; Bongards (2018), o qual implementou um sistema de controle para sistema de esgotamento sanitário combinado com drenagem urbana e verificou que para atender a demanda hidráulica do sistema, os componentes do sistema analisado foram submetidos a trabalhos intensivos, próximo da capacidade total, durante os eventos de chuva.

Esta informação é primordial para a busca pelo desenvolvimento e operação adequada neste processo, pois a manutenção e o preservação dos equipamentos envolvidos na operação de descarga das bacias são importantes para minimização de

ocorrência de extravasamentos para o meio ambiente e recuperação de óleos e graxas no através do sistema de tratamento de efluentes.

Também foram calculados também os Índices de Similaridade (IS) considerando as categorizações para os agrupamentos. Os IS, mostrados na Tabela 6, não apresentaram resultados altos, como comentado por Yu et al. (2013). Entretanto tal situação não deve ser comparada rigorosamente em função dos diferentes atributos analisados nos estudos e a relevância da opinião e experiência de quem interpretará o modelo.

Todavia, todos os IS apresentaram um valor baixo, indicando ausência ou baixa similaridade, descrevendo o baixo risco de ocorrer transbordamento dado à ausência ou ao baixo índice pluviométrico.

Tabela 6 - Índice de Similaridade (IS) de dois grupos categorizados por chuva e parâmetros de transbordamentos do sistema

	CED 1	CED 2	CETH 1	CETH 2	CETH 3
CC 1	0,000	0,232	0,000	0,232	0,000
CC 2	0,000	0,241	0,000	0,241	0,000

Fonte: O Autor, 2019.

Quanto às mudanças abruptas nas séries históricas, foram identificados ou não *change-points* em cada série temporal, sendo exibidos na Tabela 7 e Figuras 22 a 25, referentes ao índice pluviométrico e aos níveis das bacias de contenção.

Tabela 7 - Quantidade de *change-points* nas séries temporais de nível e pluviometria;

Série Histórica	Quantidade de *change-point*
CET 02	18
CET 03	12
CET 07	7
CET 09	18
OET 02	17
OET 03	7
OET 07	0
OET 09	7
Pluviométrica	13

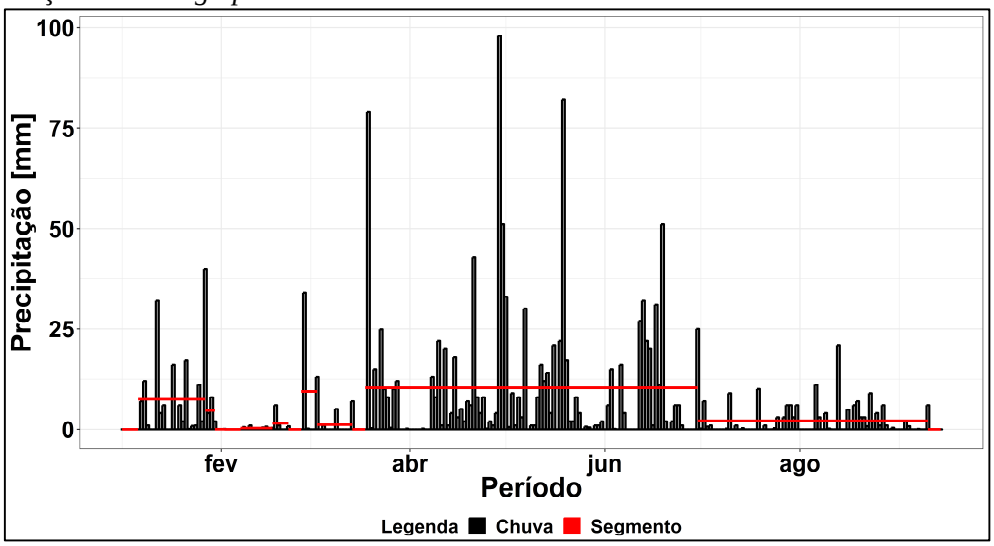
Fonte: O Autor, 2019.

A representação de cada série baseada na mediana diária do percentual de nível mostra-se uma boa apresentação quanto à possibilidade da preservação do comportamento da série, proporcionando uma série suavizada sem perda de informação quanto a tendências.

Todavia, aberrações de nível ao longo da série são omitidas, como o transbordamento pontual na OET 07 e durante alguns dias na CET 07, onde os níveis apresentados não sugerem a alusão quanto a ocorrência de transbordamento.

A ausência de um padrão universal ocorre devido às diferentes características físicas de cada bacia, bem como à topologia do sistema de drenagem, pois diferentemente de sistemas urbanos do tipo separador absoluto, redes isoladas para esgotamento sanitário de drenagem urbana, o sistema de drenagem industrial aqui analisado permite a comunicação entre os componentes do sistema através de tubovias, proporcionando influência nas séries históricas devido a operação.

Figura 22 – Série temporal de precipitação no período em análise contendo a identificação dos *change-point*.



Fonte: O Autor, 2019.

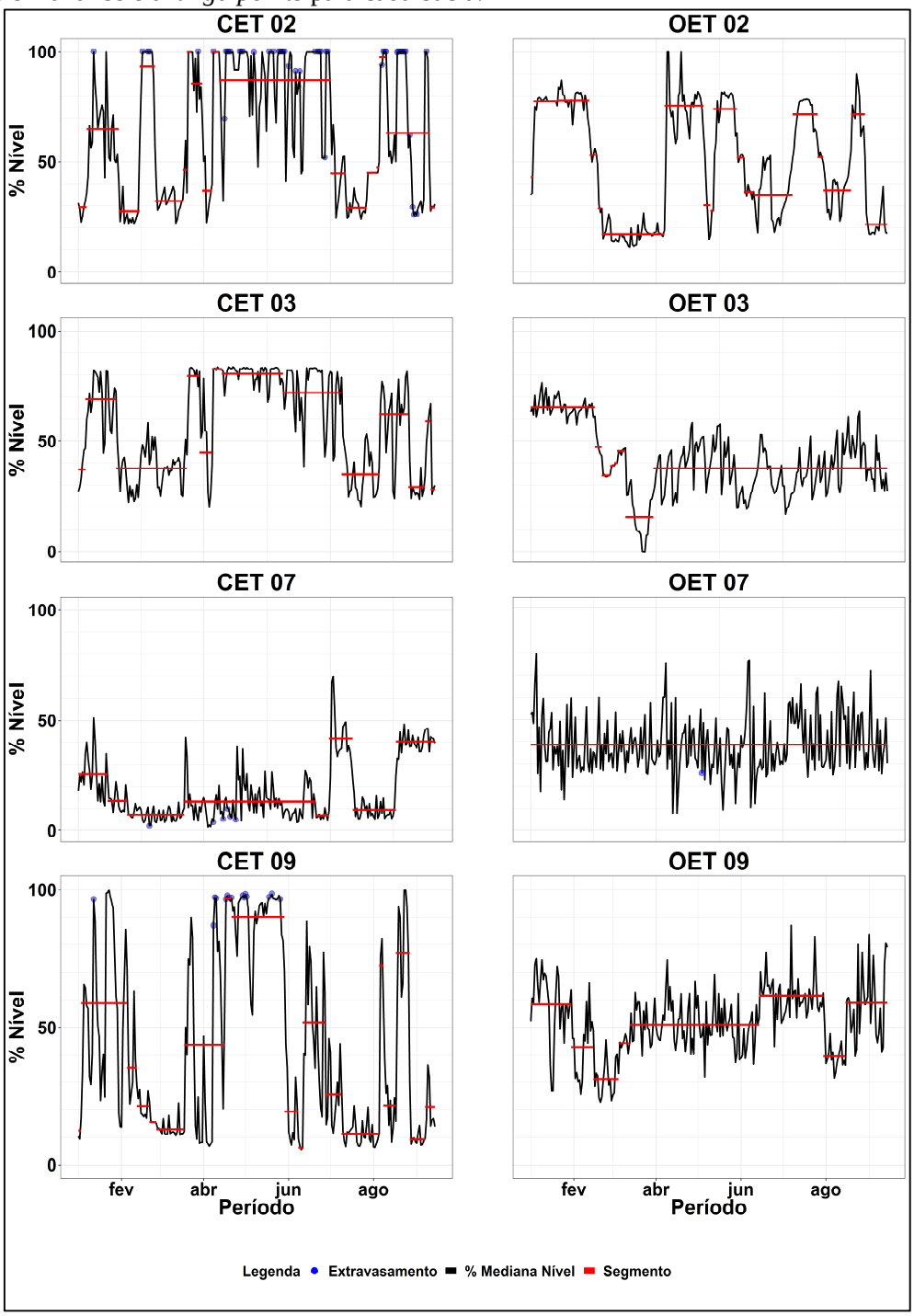
Quanto à operação, constata-se que as OET 03 e OET 09 seguem durante um longo período com uma variação de dados bem distribuída no em torno da tendência central. Esta situação caracteriza uma boa operação deste sistema analogamente a OET 07, porém com características de parâmetros estatísticos distintos, pois possivelmente os pontos de máximos locais das séries são ocasionados por interferência externa ou ocorrências processuais, e estes são bem assimilados, onde o processo de contenção e descarga de efluentes ocorre de maneira satisfatória.

A CET 03, apesar de não apresentar transbordamento, é um componente que demanda maior cuidado para o sistema, pois a mesma se mantém com média de volume próximo a 85% da capacidade de retenção durante um longo período, o que pode

impossibilitar a assimilação e armazenamento de eventuais efluentes que adentrem o sistema.

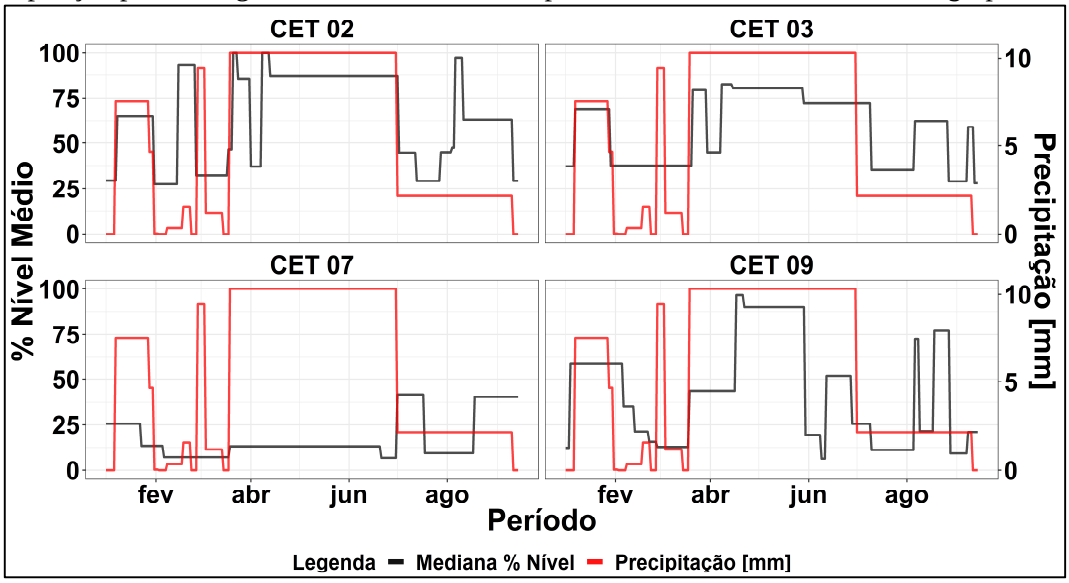
A bacia CET 07 apresenta baixo valor de mediana do nível acumulado e rápidas variações ao ponto de ocorrer transbordamento, sugerindo alta vulnerabilidade da bacia mas boa resiliência.

Figura 23 – Séries históricas referentes a mediana dos níveis de efluente acumulado nas bacias em análise e *change-points* para cada bacia.



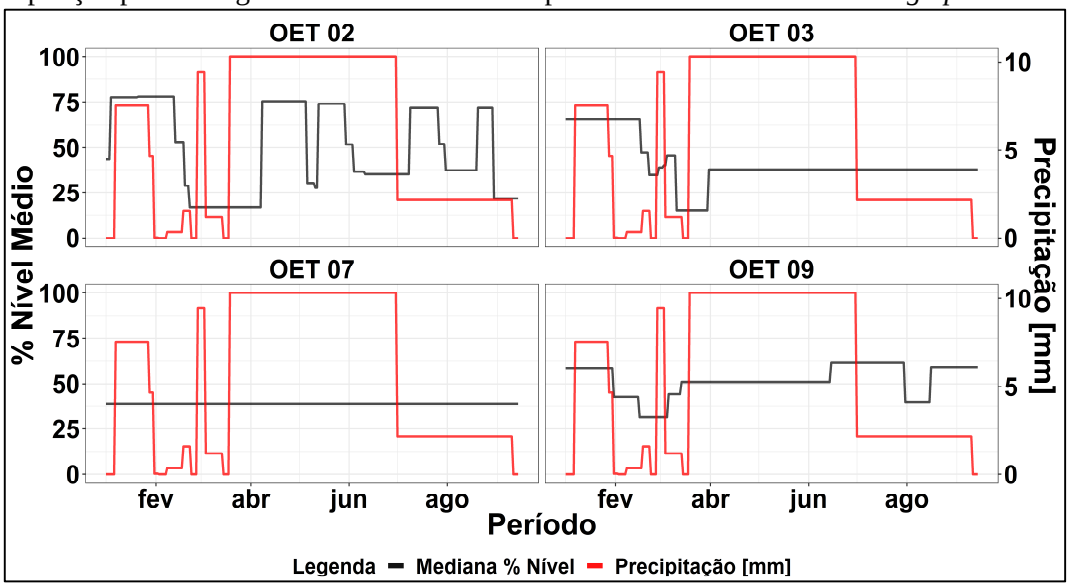
Fonte: O Autor, 2019.

Figura 24 – Médias de percentual de nível das bacias de efluente contaminado e de precipitação para os segmentos de cada série temporal em nos intervalos entre *change-points*.



Fonte: O Autor, 2019.

Figura 25 – Médias de percentual de nível das bacias de efluente oleoso e de precipitação para os segmentos de cada série temporal nos intervalos entre *change-point*.



Fonte: O Autor, 2019.

5.3 MODELAGEM PREDITIVA DE TRANSBORDAMENTO

De posse dos atributos independentes, e do atributo dependente categórico de interesse, foram construídos 24 modelos preditivos utilizando *Random Forest* e KNN. Para cada modelo foi gerado uma matriz de confusão, conforme a Tabela 1, com *scores* de verdadeiro positivo, verdadeiro negativo, falso positivo e falso negativo, que podem ser visualizadas no material no Apêndice.

No primeiro momento, devido ao desbalanceamento de informação (14,4% Extravasando *versus* 85,6% Operação Normal), as principais métricas para avaliação inicial foram *Kappa*, sensibilidade e acurácia. A escolha da métrica *Kappa* é importante porque apresenta ponderação sobre todos os scores da matriz de confusão.

Como definido na metodologia, três cenários foram definidos de acordo com o conjunto de variáveis empregadas para a construção de cada modelo.

Para os modelos construídos no cenário 1 os melhores resultados foram para os modelos com o algoritmo KNN forma com as técnicas de reamostragem *undersampling* e ROSE. Ambos tiveram verdadeiro positivo igual a 9, e ausência de *score* para falso negativo, ou seja, todas as predições de extravasamento foram realizadas totalmente de acordo com o ocorrido histórico. Entretanto o modelo com ROSE teve também um menor score de falso positivo, mas possui elevado número ótimo de vizinhos (k=15), refletindo a complexidade do modelo e um possível *overfitting* dado a elevada acurácia.

Dos modelos *Random Forest* no cenário 1, destacam-se os modelos sem emprego de técnica de reamostragem e com *oversampling*. Ambos têm bom desempenho para o índice *Kappa* (aproximadamente igual a 0,75) e acurácia de 0,93, e possuem, respectivamente, o bom score com falso negativo igual a 2 e 1. Os modelos com técnicas de undersampling e ROSE predizem de maneira assertiva todas as observações de extravasamento, mas contam com demasiado falso negativo.

No cenário 2 com KNN os modelos apesar de apresentarem excelentes resultados de acurácia, aproximadamente 0,90, possuem baixos valores de sensibilidade. Esta situação reflete um alto score de falso negativo, o que é extremamente indesejável no contexto de predição de extravasamento.

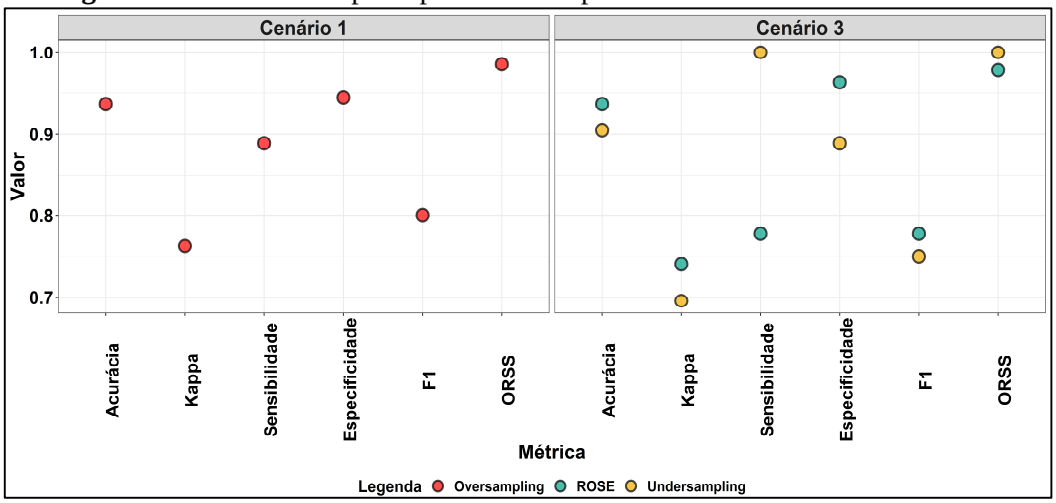
Como os modelos *Random Forest* no cenário 2 ocorre algo similar para o modelo sem emprego de técnica de reamostragem e o modelo com oversampling, que é concretizado pelos módicos valores de *Kappa* (0,57 e 0,56). O modelo com undersampling e ROSE tem bom acurácia e sensitividade, mas aponta um desequilíbrio com valores de *Kappa* menor que 0,60.

Com o conjunto de variáveis no cenário 3 utilizando KNN apenas o modelo com *undersampling* apresentou índices satisfatórios, entretanto o elevado número de vizinhos, k=24, torna o modelo muito complexo. Diferentemente do que acontece com *Random Forest* com *undersampling* e ROSE, que apresentaram bons resultados, bem equilibrados.

O modelo *Random Forest* com *undersampling* se destaca também por apresentar *mtry* = 2.

Na Figura 26 são apresentados os resultados para os modelos construídos com resultados de maior destaque, destacando que todos os modelos com os melhores resultados são baseados no algoritmo *Random Forest* e com emprego de técnica de reamostragem, ratificando a informação de melhor desempenho da *Random Forest* frente ao KNN (JAMES et al., 2013).

Figura 26 – Análise dos principais modelos preditores construídos.



Fonte: O Autor, 2019.

Os modelos escolhidos apresentam bons resultados com elevados valores para as métricas analisadas, todos com acurácia maior que 0,90 e bom equilíbrio entre as demais métricas. Entretanto, o modelo com *Random Forest* com técnica de undersampling com as variáveis do cenário 3 apresenta sensibilidade igual a 1, ou seja, ele acerta todos os momentos que ocorre o fenômeno de extravasamento na CET 09, como pode ser visto na Tabela 8.

Tabela 8 - Matriz de confusão com Random Forest undersampling no cenário 3.

Evento Predito	Evento Observado	
	Extravasamento	Operação Normal
Extravasamento	9	6
Operação Normal	0	48

Fonte: O Autor, 2019.

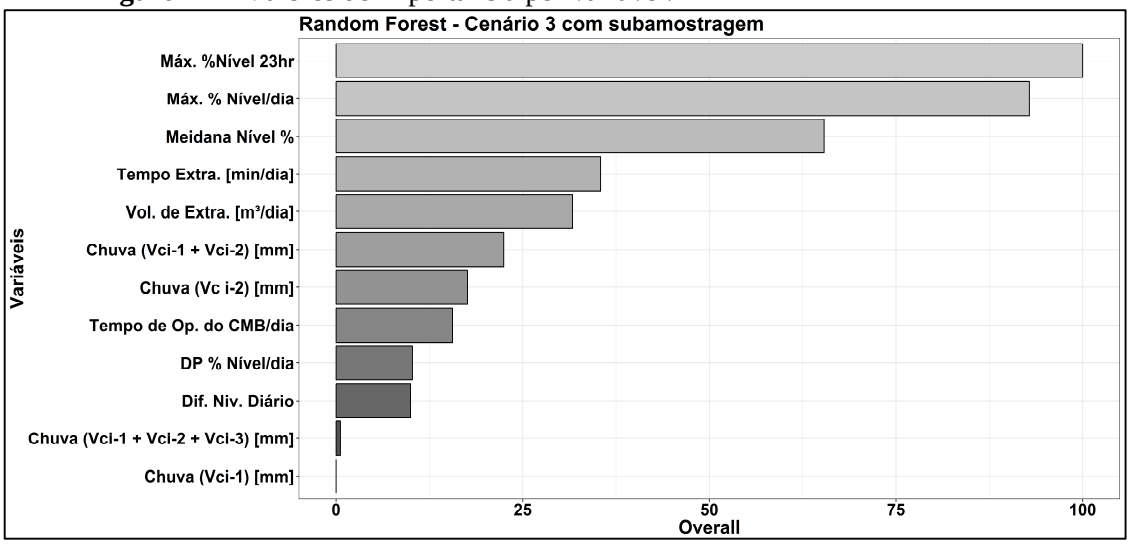
É importante garantir valores baixos ou a ausência de falsos negativos na modelagem do extravasamento. A falha na identificação do risco de extravasamento pode trazer agravantes para a empresas e para o meio ambiente. Esta ação visa proporcionar

uma boa ferramenta preditiva que forneça suporte à tomada de decisão aos profissionais envolvidos na operação do sistema de drenagem industrial.

Observou-se que o modelo *Random Forest* discutido na Tabela 8 classificou corretamente os fenômenos de extravasamento com boa distinção de classes, promovendo verdadeiros positivos com valores de probabilidade entre 0,70 e 1,0, revelando uma boa segmentação, como também pode ser visualizado em Apêndice, apresentando a estimativa da probabilidade de ocorrer o extravasamento ao longo de todo o período analisado.

A variável mais importante para o modelo preditivo escolhido é o máximo nível percentual diário da CET 09 no intervalo de 23:00 às 23:59, como pode ser visualizado na Figura 27 baseado no Índice Gini.

Figura 27 – Valores de importância por variável.



Fonte: O Autor, 2019.

Os baixos valores de importância para chuva diária (Vc_{i-1}) acumulada [mm/dia] no dia anterior de refletem a necessidade de um monitoramento mais detalhado desse fenômeno, pois, possivelmente, a informação do evento hidrológico está subdividia com as próprias variáveis relacionadas a precipitação pluviométrica diária acumulada.

As variáveis mais importantes ratificam que o sistema transborda devido aos altos níveis de efluente acumulado. Contudo, questões operacionais não foram verificadas para maior conhecimento, como a conformidade dos padrões de qualidade do efluente para transporte a jusante e tratamento na Estação de Tratamento de Efluentes Industriais.

Observou-se também que o número de árvores necessário para *Random Forest* pode ser reduzido em um valor próximo a um quinto das 1500 árvores de decisão iniciais, devido a estabilização do erro OOB, como pode ser visualizado em Apêndice.

6 CONCLUSÃO

No presente trabalho foi realizado mineração de dados para melhor entendimento sobre a operação do sistema de coleta e retenção de efluentes industriais em um refinaria de petróleo, sendo proposto um modelo de predição de transbordamento de efluentes em uma bacia, CET 09.

Verificou-se que o conjunto de metodologias de aprendizagem não-supervisionada aqui utilizadas fornecem meios para a extração de informações quanto a eventos hidroclimáticos e processuais em cenários com pouca disponibilidade de dados e sem necessidade de emprego de técnicas auxiliares, como simulações hidrodinâmicas e inserção de novos instrumentos de medição para obtenção de mais dados e informações.

A análise de agrupamentos permitiu melhor sensibilidade sobre a complexidade quanto à análise do comportamento do sistema de contenção de efluentes, ratificando que os períodos com elevada média de tempo de trabalho das estações elevatórias norteiam a necessidade do uso máximo do sistema para a ausência de transbordamento. Podendo ser interpretada para desenvolvimento de planos de manutenção.

Foi identificado que na ausência de precipitação ou ocorrência de baixos volumes diários de precipitação existem transbordamentos, e o percentual em períodos de chuva é maior que o valor natural esperado de falhas, tais situações são consideradas anômalas e evidenciam falhas operacionais.

Compreende-se também que os valores do Índice de Similaridade (IS) dependem da qualidade dos dados disponíveis e da interpretação do pesquisador, e podem ser interpretados de modo que o pesquisador ou operador entender que seja coerente e tangível, não sendo necessário apenas um valor do Índice de Similaridade como referência, como sugerido por (YU et al., 2013).

A divisão de agrupamentos baseada em heurística a partir da visualização do dendrograma não proporciona ganho de informação quanto ao Índice de Similaridade. Contudo um maior número de grupos proporciona maior sensibilidade para a análise do problema.

De acordo com os *change-points* foi possível constatar que os períodos com ocorrência de transbordamento coincidem esperadamente com os períodos chuvosos, onde há aumento de média. Foi também notado que o subsistema de efluentes

contaminados, comparado ao oleoso, é mais vulnerável a ocorrência de chuvas, dominando a frequência de transbordamentos.

Mesmo com falta de informação ou dados censurados é perceptível a vantagem e possibilidade de aplicação da análise de *change-point*, inclusive para análise de padrões e inferências em sistemas de drenagem urbana.

O sistema de contenção de efluentes industriais aqui analisado não foi capaz de cumprir completamente com sua função de retenção de volumes extras proporcionados por eventos externos, como a ocorrência de precipitação pluviométrica. Logo, recomenda-se reavaliar as conformidades operacionais e o risco associado à probabilidade de falha, transbordamento, estimada para o projeto do sistema para projetos futuros e possíveis melhorias no sistema em operação. Este fato se torna ainda mais relevante se considerado cenários com possíveis mudanças climáticas.

A complexidade e a não linearidade que envolve o fenômeno de transbordamento, e a ausência de dados relacionados à precipitação pluviométrica em menor escala de tempo tornam a modelagem empírica deste fenômeno um desafio. Não obstante, foi realizada a modelagem com horizonte preditivo para o dia posterior quanto à classificação se a bacia de contenção de efluentes em uma refinaria, CET 09, transbordaria ou não.

Os melhores resultados de modelos construídos para a modelagem sobre a CET 09 foram obtidos a partir do algoritmo *Random Forest* com emprego de técnica de *oversampling*, *undersampling* e ROSE. O melhor resultado foi para o conjunto de variáveis do cenário 3 com o emprego de técnica de reamostragem do tipo *undersampling*.

Consta-se que o trabalho desenvolvido pode ser aplicado, em situações análoga de dados e informação, para sistemas de drenagem urbana. Podendo desta maneira contribuir para planos de saneamento, estudos de manejo de águas pluviais e operação de tanques de retenção em zona urbana, conhecidos como “piscinões”.

Contudo, a ausência de equação matemática com parâmetros definidos, a impossibilidade de predição para menores intervalos de tempo, a ausência de características operacionais sobre a qualidade do efluente, a incapacidade de ponderar a espacialização de chuvas e implicações potenciais a jusante do sistema de drenagem industrial são desvantagens do estudo realizado.

7 SUGESTÃO PARA TRABALHOS FUTUROS

Como sugestão para trabalhos futuros, tem-se:

1) Análise de incerteza do modelo de predição;
2) Aplicação do melhor modelo preditivo para outro conjunto de dados da CET 09;
3) Predição de transbordamento em abordagem considerando o sistema;
4) Desenvolvimento de controle preditivo dos tanques de contenção de efluentes;
5) Modelagem preditiva de transbordamento considerando transientes hidráulico e capacidade hidrodinâmica do sistema;
6) Análise da influência da qualidade do efluente quanto à operação do sistema e na modelagem do sistema.

REFERÊNCIAS

ALMEIDA, L. T. DE. **ESPACIALIZAÇÃO DE CHUVAS INTENSAS: UMA NOVA PROPOSTA**. Dissertação de Mestrado. Programa de Pós-Graduação em Meteorologia Aplicada, Universidade Federal de Viçosa, Viçosa-MG, 2017.

ARTINA, S. et al. Simulation of a storm sewer network in industrial area: Comparison between models calibrated through experimental data. **Environmental Modelling & Software**, v. 22, n. 8, p. 1221–1228, ago. 2007.

BALL, J. E.; LUK, K. C. Modeling Spatial Variability of Rainfall over a Catchment. **Journal of Hydrologic Engineering**, v. 3, n. 2, p. 122–130, abr. 1998.

BARRETO, T. B. et al. **Comparação entre metodologias de classificação da condição hídrica anual da bacia hidrográfica**. Florianópolis: XX Simpósio Brasileiro de Recursos Hídricos., 2017. Disponível em: <http://evolvedoc.com.br/xxiisbrh/detalhes-657_comparacao-entre-metodologias-de-classificacao-da-condicao-hidrica-anual-da-bacia-hidrografica>.

BASISTHA, A.; ARYA, D. S.; GOEL, N. K. Spatial Distribution of Rainfall in Indian Himalayas – A Case Study of Uttarakhand Region. **Water Resources Management**, v. 22, n. 10, p. 1325–1346, 18 out. 2008.

BASTOS, P. C. **EFEITOS DA URBANIZAÇÃO SOBRE VAZÕES DE PICO DE ENCHENTE**. Vitória, ES: Programa de Pós- Graduação em Engenharia Ambiental do Centro Tecnológica da Universidade Federal do Espírito Santo, 2009.

BEAULIEU, C.; CHEN, J.; SARMIENTO, J. L. Change-point analysis as a tool to detect abrupt climate variations. **Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences**, v. 370, n. 1962, p. 1228–1249, 2012.

BEN-DAVID, S.; PÁL, D.; SIMON, H. U. Stability of k-Means Clustering. In: **Learning Theory**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007. p. 20–34.

BENGIO, Y.; GRANDVALET, Y. No Unbiased Estimator of the Variance of K-Fold Cross-Validation. **Journal of Machine Learning Research**, v. 5, p. 1089–1105, 2003.

BERSINGER, T. et al. Online monitoring and conditional regression tree test: Useful tools for a better understanding of combined sewer network behavior. **Science of The Total Environment**, v. 625, p. 336–343, jun. 2018.

BRASIL. **Manual de hidrologia básica para estruturas de drenagem**. 2. ed ed. Rio de Janeiro: Departamento Nacional de Infra-Estrutura de Transportes - DNIT, 2005.

BREIMAN, L. et al. **Classification and Regression Trees**. Boca Raton, FL: Chapman and Hall/CR, 1984.

BREIMAN, L. RANDOM FORESTS Leo. **Machine Learning**, v. 45, n. 1, p. 5–32, 2001.

BROCK, G. et al. clValid : An R Package for Cluster Validation. **Journal of Statistical Software**, v. 25, n. 4, p. 1–22, 2008.

CARVALHO, D. F. DE; SILVA, L. D. B. DA. **Hidrologia**. Rio de Janeiro: UFRRJ, 2006.

CEMBRANO, G. et al. Optimal control of urban drainage systems . A case study. **Control Engineering Practice**, v. 12, n. 1, p. 1–9, 2004.

CHEN, F. W.; LIU, C. W. Estimation of the spatial rainfall distribution using inverse

distance weighting (IDW) in the middle of Taiwan. **Paddy and Water Environment**, v. 10, n. 3, p. 209–222, 2012.

CHEN, G.; GE, Z. SVM-tree and SVM-forest algorithms for imbalanced fault classification in industrial processes. **IFAC Journal of Systems and Control**, v. 8, p. 100052, jun. 2019.

CHEN, T. et al. Comparison of Spatial Interpolation Schemes for Rainfall Data and Application in Hydrological Modeling. **Water**, v. 9, n. 5, p. 342, 11 maio 2017.

CLIMATEMPO. **Climatologia: São Francisco do Conde/BA**. Disponível em: <https://www.climatempo.com.br/climatologia/4832/saofranciscodoconde-ba>. Acesso em: 6 set. 2018.

COHEN, J. A Coefficient of Agreement for Nominal Scales. **Educational and Psychological Measurement**, v. 20, n. 1, p. 37–46, 2 abr. 1960.

COLLISCHONN, W.; TASSI, R. **Introduzindo Hidrologia**. Porto Alegre: IPH UFRGS, 2008.

CONAMA - CONSRLHO NACIONAL DO MEIO AMBIENTE. **RESOLUÇÃO Nº 20.** Ministério do Meio Ambiente. Brasil, 1986. Disponível em: <http://www2.mma.gov.br/port/conama/res/res86/res2086.html>

COSTA, M.; GONÇALVES, A. M.; TEIXEIRA, L. Change-point detection in environmental time series based on the informational approach. **Electronic Journal of Applied Statistical Analysis**, v. 9, n. 2, p. 267–296, 2016.

DIETTERICH, T. G. An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees. **Machine Learning**, v. 40, p. 139–157, 2000a.

DIETTERICH, T. G. Ensemble Methods in Machine Learning. Lecture Notes in Computer Science, vol 1857, Berlim, 2000.

DIRKS, K. N. et al. High-resolution studies of rainfall on Norfolk Island. **Journal of Hydrology**, v. 208, n. 3–4, p. 187–193, jul. 1998.

DIYA'UDDEEN, B. H.; DAUD, W. M. A. W.; ABDUL, A. A. R. Treatment technologies for petroleum refinery effluents: A review. **Process Safety and Environmental Protection**, v. 89, n. 2, p. 95–105, 2011.

DOURADO, C. DA S. **Mineração de dados climáticos para análise de eventos extremos de precipitação**. Dissertação de Mestrado. Programa de Pós-Graduação em Engenharia Agrícola da Faculdade de Engenharia Agrícola. UNICAMP, 2013.

EPA. **Report to Congress on Impacts and Control of Combined Sewer Overflows and Sanitary Sewer Overflows Fact Sheet**. Washington, D.C.: U.S. Environmental Protection Agency, 2004.

FEELDERS, A.; DANIELS, H.; HOLSHEIMER, M. Methodological and practical aspects of data mining. **Information & Management**, v. 37, n. 5, p. 271–281, ago. 2000.

GE, Z. et al. Data Mining and Analytics in the Process Industry: The Role of Machine Learning. **IEEE Access**, v. 5, p. 20590–20616, 2017.

GONZALEZ, C. C. C. C. et al. **Aplicação de Modelo Hidrológico Dinâmico ao Sistema de Coleta e Transporte de Efluentes de uma Refinaria de Petróleo**. Salvador: JESAM - Jornada de Engenharia Sanitária e Ambiental, 2013

GOUTTE, C.; GAUSSIER, E. **A Probabilistic Interpretation of Precision, Recall and**

F-Score, with Implication for Evaluation. 2005.

HANDL, J.; KNOWLES, J.; KELL, D. B. Computational cluster validation in post-genomic data analysis. **Bioinformatics**, v. 21, n. 15, p. 3201–3212, 1 ago. 2005.

HÄRDLE, W.; HLÁVKA, Z.; KLINKE, S. **XploRe® - Application Guide**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2000.

HARVEY, R. R.; MCBEAN, E. A. Predicting the structural condition of individual sanitary sewer pipes with random forests. **Canadian Journal of Civil Engineering**, v. 41, n. 4, p. 294–303, abr. 2014.

HASAN, M. M.; CROKE, B. F. W. **Filling gaps in daily rainfall data: a statistical approach**. Adelaide, Austrália: 20th International Congress on Modelling and Simulation, 2013Disponível em: <http://www.mssanz.org.au/modsim2013/A9/hasan.pdf>

HASSANAT, A. B.; ABBADI, M. A.; ALTARAWNEH, G. A. Optimal K parameter for KNN Classifier with square root. **International Journal of Computer Science and Information Security**, v. 12, n. 8, p. 33–39, 2014.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning**. New York, NY: Springer New York, 2009.

HE, H.; GARCIA, E. A. Learning from Imbalanced Data. **IEEE Transactions on Knowledge and Data Engineering**, v. 21, n. 9, p. 1263–1284, set. 2009.

HENNIG, C. Cluster-wise assessment of cluster stability. **Computational Statistics & Data Analysis**, v. 52, n. 1, p. 258–271, set. 2007.

HODGSON, J. E.; BENDIAK, L. C. Stormwater Management for Petroleum Refineries. **Canadian Water Resources Journal**, v. 12, n. 3, p. 38–47, jan. 1987.

HOSSIN, M. B.; SULAIMAN, M. N. A Review on Evaluation Metrics for Data Classification Evaluations. **International Journal of Data Mining & Knowledge Management Process**, v. 5, n. 2, p. 01-11, 2015.

HU, W.; CHEN, T.; SHAH, S. L. Detection of Frequent Alarm Patterns in Industrial Alarm Floods Using Itemset Mining Methods. **IEEE Transactions on Industrial Electronics**, v. 65, n. 9, p. 7290–7300, 2018.

JAMES, G. et al. **An Introduction to Statistical Learning**. New York, NY: Springer New York, 2013. v. 103

JOLIFFE, I; STEPHENSON, D. **Forecast verification : a practitioner's guide in atmospheric science**. 2nd ed. ISBN-10: 0470660716. John Wiley & Sons, 2012.

JYOTI, K.; SINGH, S. Data Clustering Approach to Industrial Process Monitoring , Fault Detection and Isolation. **International Journal of Computer Applications**, v. 17, n. 2, p. 41–45, 2011.

KAMBLE, S. S.; GUNASEKARAN, A.; GAWANKAR, S. A. Sustainable Industry 4.0 framework: A systematic literature review identifying the current trends and future perspectives. **Process Safety and Environmental Protection**, v. 117, p. 408–425, 2018.

KASSAMBARA, A. **Practical Guide To Cluster Analysis in R: Unsupervised Machine Learning**. [s.l.] STHDA, 2017.

KAUFMAN, L.; ROUSSEEUW, P. Partitioning Around Medoids (Program PAM). In: Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley & Sons, 1990.

p. 68–125.

KILLICK, R.; ECKLEY, I. A. changepoint: An R Package for changepoint analysis. **Journal of Statistical Software**, v. 58, n. 3, p. 1–19, 2014.

KIM, J.-H. Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. **Computational Statistics & Data Analysis**, v. 53, n. 11, p. 3735–3745, set. 2009.

KRAWCZYK, B. Learning from imbalanced data: open challenges and future directions. **Progress in Artificial Intelligence**, v. 5, n. 4, p. 221–232, 22 nov. 2016.

KUHN, M. **caret: Classification and Regression Training**R package version 6.0-81, , 2018. Disponível em: <https://cran.r-project.org/package=caret>

KUHN, M.; JOHNSON, K. **Applied Predictive Modeling**. New York, NY: Springer New York, 2013.

KUSCH, W.; HAAG, T.; BONGARDS, M. **Control of distributed tanks for stormwater treatment in combination with a wastewater treatment plant and a sewer system**. 7th International Energy and Sustainability Conference (IESC). **Anais**...Colônia, Alemanha: IEEE, 2018Disponível em: <https://ieeexplore.ieee.org/document/8439994>

LEITÃO, G.; AFFONSO GUEDES, L. Real Time Alarm Processing for Predictive Failure Diagnosis in Petrochemical Plants. **IEEE Latin America Transactions**, v. 14, n. 7, p. 3481–3489, 2016.

LIN, W. C. et al. Clustering-based undersampling in class-imbalanced data. **Information Sciences**, v. 409–410, p. 17–26, 2017.

LIU, L. et al. Spatial fuzzy clustering approach to characterize flood risk in urban storm water drainage systems. **Natural Hazards**, v. 83, n. 3, p. 1469–1483, 2016.

LIU, Y. et al. **Understanding of Internal Clustering Validation Measures**. 2010 IEEE International Conference on Data Mining. **Anais**...IEEE, dez. 2010Disponível em: <http://ieeexplore.ieee.org/document/5694060/>

LÖWE, R.; MADSEN, H.; MCSHARRY, P. Objective Classification of Rainfall in Northern Europe for Online Operation of Urban Water Systems Based on Clustering Techniques. **Water**, v. 8, n. 3, p. 87, 2016.

MAILHOT, A.; TALBOT, G.; LAVALLÉE, B. Relationships between rainfall and Combined Sewer Overflow (CSO) occurrences. **Journal of Hydrology**, v. 523, p. 602–609, 2015.

MARCUS, A. C. .; EKPETE, O. A. Impact of Discharged Process Wastewater from an Oil Refinery on the Physicochemical Quality of a Receiving Waterbody in Rivers State, Nigeria. **IOSR Journal of Applied Chemistry (IOSR-JAC)**, v. 7, n. 12, p. 01–08, 2014.

MARTINO, G. DE et al. Pollution Reduction in Receivers: Storm-Water Tanks. **Journal of Urban Planning and Development**, v. 137, n. 1, p. 29–38, mar. 2011.

MENARDI, G.; TORELLI, N. Training and assessing classification rules with imbalanced data. **Data Mining and Knowledge Discovery**, v. 28, n. 1, p. 92–122, 30 jan. 2014.

MIRÁS-AVALOS, J. M. et al. Mapping monthly rainfall data in Galicia (NW Spain) using inverse distances and geostatistical methods. **Advances in Geosciences**, v. 10, p.

51–57, 26 abr. 2007.

MONTSERRAT, A. et al. Using data from monitoring combined sewer overflows to assess, improve, and maintain combined sewer systems. **Science of the Total Environment**, v. 505, p. 1053–1061, 2015.

MOUNCE, S. R. et al. Predicting combined sewer overflows chamber depth using artificial neural networks with rainfall radar data. **Water Science and Technology**, v. 69, n. 6, p. 1326–1333, mar. 2014.

OCAMPO-MARTINEZ, C. **Model Predictive Control of Wastewater Systems**. 1. ed. Springer-Verlag London, 2010.

OLIVEIRA-ESQUERRE, K. P. et al. Taking advantage of storm and waste water retention basins as part of water use minimization in industrial sites. **Resources, Conservation and Recycling**, v. 55, n. 3, p. 316–324, 2011.

OSIN, O. A.; YU, T.; LIN, S. Oil refinery wastewater treatment in the Niger Delta, Nigeria: current practices, challenges, and recommendations. **Environmental Science and Pollution Research**, v. 24, n. 28, p. 22730–22740, 2017.

PAZ, A. R. **Hidrologia Aplicada**. Caxias do Sul (RS): UERGS, 2004.

PEBESMA, E. **The meuse data set: a brief tutorial for the gstat R package.**ViennaR Foundation for Statistical Computing, , 2018. Disponível em: <https://cran.r-project.org/web/packages/gstat/vignettes/gstat.pdf>

PELLICONE, G. et al. Application of several spatial interpolation techniques to monthly rainfall data in the Calabria region (southern Italy). **International Journal of Climatology**, n. April, p. 3651–3666, 2018.

PESSANHA, J. F. M. et al. CONSTRUINDO TIPOLOGIAS DE CURVAS DE CARGA COM O PROGRAMA R. **Revista Eletrônica Pesquisa Operacional para o Desenvolvimento**, v. 7, p. 29–54, 2015.

PIECH, C. **K Means**. Disponível em: <http://stanford.edu/~cpiech/cs221/handouts/kmeans.html>. Acesso em: 5 maio. 2019.

PIHUR, V.; BROCK, G. N.; DATTA, S. Cluster Validation for Microarray Data: An Appraisal. In: **Advances in Multivariate Statistical Analysis**. [s.l: s.n.]. p. 79–94.

PIOT, M. Clustering, Distance Methods and Ordination. In: **Statistics in Climate Sciences**. Berna, Suiça: University of Bern, 2014.

R CORE TEAM. **R: A Language and Environment for Statistical Computing**Vienna, AustriaR Foundation for Statistical Computing, , 2018. Disponível em: <https://www.r-project.org>

RAMOS, A. M.; SANTOS, L. A. R. DOS; FORTES, L. T. G. **Normais climatológicas do Brasil 1961-1990**. Brasília, DF: INMET, 2009.

ROSEN, C.; RÖTTORP, J.; JEPPSSON, U. Multivariate on-line monitoring: challenges and solutions for modern wastewater treatment operation. **Water Science and Technology**, v. 47, n. 2, p. 171–179, jan. 2003.

SANDOVAL, S. et al. The evaluation of rainfall influence on combined sewer overflows characteristics: the Berlin case study. **Water Science and Technology**, v. 68, n. 12, p. 2683–2690, dez. 2013.

SANTANA, S. P. B.; PESSOA, R. W. S.; OLIVEIRA-ESQUERRE, K. P. **Reliability**

analysis of a containment system, transport and segregation of effluents. Society For Risk Analysis Annual Meeting 2017. **Anais**...Arlignton, VA: Society for Risk Analysis (SRA), 2017Disponível em: <http://birenheide.com/sra/2017AM/program/singlesession.php3?sessid=P>

SAUCEDO-MARTÍNEZ, J. A. et al. Industry 4.0 framework for management and operations: a review. **Journal of Ambient Intelligence and Humanized Computing**, v. 9, n. 3, p. 789–801, 2018.

SCHEIDT, F. A.; ANGELICA, M.; BRUNETTO, D. C. **Modelagem Chuva-vazão utilizando Redes Neurais Artificiais e Algoritmos Genéticos**. XXXI Congresso da Sociedade Brasileira de Computação. **Anais**...Natal, RN: Sociedade Brasileira de Computação, 2011Disponível em: <http://dimap.ufrn.br/csbc2011/anais/eventos/contents/WCAMA/Wcama_Sessao1_Artigo1_Scheidt.pdf>

SCHMITT, T. G.; THOMAS, M.; ETTRICH, N. Analysis and modeling of flooding in urban drainage systems. **Journal of Hydrology**, v. 299, n. 3–4, p. 300–311, 2004.

SCHOLZ, M. Classification of flood retention basins: The Kaiserstuhl case study. **Environmental and Engineering Geoscience**, v. 14, n. 2, p. 61–80, 2008.

SCHROEDER, K. et al. Evaluation of effectiveness of combined sewer overflow control measures by operational data. **Water Science & Technology**, v. 63, n. 2, p. 325, 2011.

SHALABI, L. AL; SHAABAN, Z.; KASASBEH, B. Data Mining: A Preprocessing Engine. **Journal of Computer Science**, v. 2, n. 9, p. 735–739, 1 set. 2006.

SILVA, L. M. O. DA. **Uma Aplicação de Árvores de Decisão, Redes Neurais e KNN para a Identificação de Modelos ARMA Não-Sazonais e Sazonais**. Tese de Doutorado. Programa de Pós-graduação em Engenharia Elétrica. PUC Rio, Rio de Janeiro, 2005.

SILVA, L. A.; PERES, S. M.; BOSCARIOLI, C. **Introdução à Mineração de Dados Com Aplicações em R**. 1. ed. ed. Rio de Janeiro: Elsevier, 2016.

SYAFRUDIN, M. et al. Performance Analysis of IoT-Based Sensor, Big Data Processing, and Machine Learning Model for Real-Time Monitoring System in Automotive Manufacturing. **Sensors**, v. 18, n. 9, p. 2946, 4 set. 2018.

TANTITHAMTHAVORN, C.; HASSAN, A. E.; MATSUMOTO, K. The Impact of Class Rebalancing Techniques on the Performance and Interpretation of Defect Prediction Models. **IEEE Transactions on Software Engineering**, p. 1–20, 2018.

TATSCH, J. D. **inmetr R package (v 0.2.5): Historical Data from Brazilian Meteorological Stations in R.**Santa Maria.Zenodo, 2018. Disponível em: <https://github.com/lhmet/inmetr>

TEIXEIRA, L. M. L. **Análise de Change-points em Séries Temporais**. Dissertação de Mestrado. Estatística.Universidade do Minho, 2012.

THOMAS, M. C.; ZHU, W.; ROMAGNOLI, J. A. Data mining and clustering in chemical process databases for monitoring and knowledge discovery. **Journal of Process Control**, v. 67, p. 160–175, jul. 2018.

THORNDAHL, S.; SCHAARUP-JENSEN, K.; JENSEN, J. B. Probabilistic modelling of combined sewer overflow using the First Order Reliability Method. **Water Science & Technology**, v. 57, n. 9, p. 1337–1344, 2008.

TODESCHINI, S.; PAPIRI, S.; CIAPONI, C. Performance of stormwater detention tanks

for urban drainage systems in northern Italy. **Journal of Environmental Management**, v. 101, n. December 2003, p. 33–45, jun. 2012.

TOMAZ, P. Capítulo 3 Período de retorno. In: **Cálculos hidrológicos e hidráulicos para obras municipais**. [s.l: s.n.]. p. 65–73.

VENKATASUBRAMANIAN, V.; RENGASWAMY, R.; YIN, K. A re v iew of process fault detection and diagnosis Part I : Quantitati v e model-based methods. v. 27, p. 293–311, 2003.

WANG, H.; KILLICK, R.; FU, X. Distributional change of monthly precipitation due to climate change: Comprehensive examination of dataset in southeastern United States. **Hydrological Processes**, v. 28, n. 20, p. 5212–5219, 2014.

WARD, J. H. Hierarchical Grouping to Optimize an Objective Function. **Journal of the American Statistical Association**, v. 58, n. 301, p. 236, mar. 1963.

WEATHER. **Water Cycle Poster**. Disponível em: <https://www.weather.gov/jetstream/hydrocycle_max>. Acesso em: 22 jul. 2019.

WEESE, M. et al. Statistical Learning Methods Applied to Process Monitoring: An Overview and Perspective. **Journal of Quality Technology**, v. 48, n. 1, p. 4–24, 21 jan. 2016.

WITTEN, I. H. et al. **Data Mining Practical Machine Learning Tools and Techniques**. Fourth Edi ed. Cambridge, MA: Elsevier, 2017.

XU, S. et al. Data cleaning in the process industries. **Reviews in Chemical Engineering**, v. 31, n. 5, 1 jan. 2015.

YAZDI, J.; CHOI, H. S.; KIM, J. H. A methodology for optimal operation of pumping stations in urban drainage systems. **Journal of Hydro-environment Research**, v. 11, p. 101–112, jun. 2016.

YU, Y. et al. Cluster analysis for characterization of rainfalls and CSO behaviours in an urban drainage area of Tokyo. **Water Science and Technology**, v. 68, n. 3, p. 544–551, 2013.

YU, Y. et al. Simple Method for Calculating Hydraulic Behavior of Combined Sewer Overflow from Rainfall Event Data. **Journal of Water Resources Planning and Management**, v. 144, n. 10, p. 04018061, 2018.

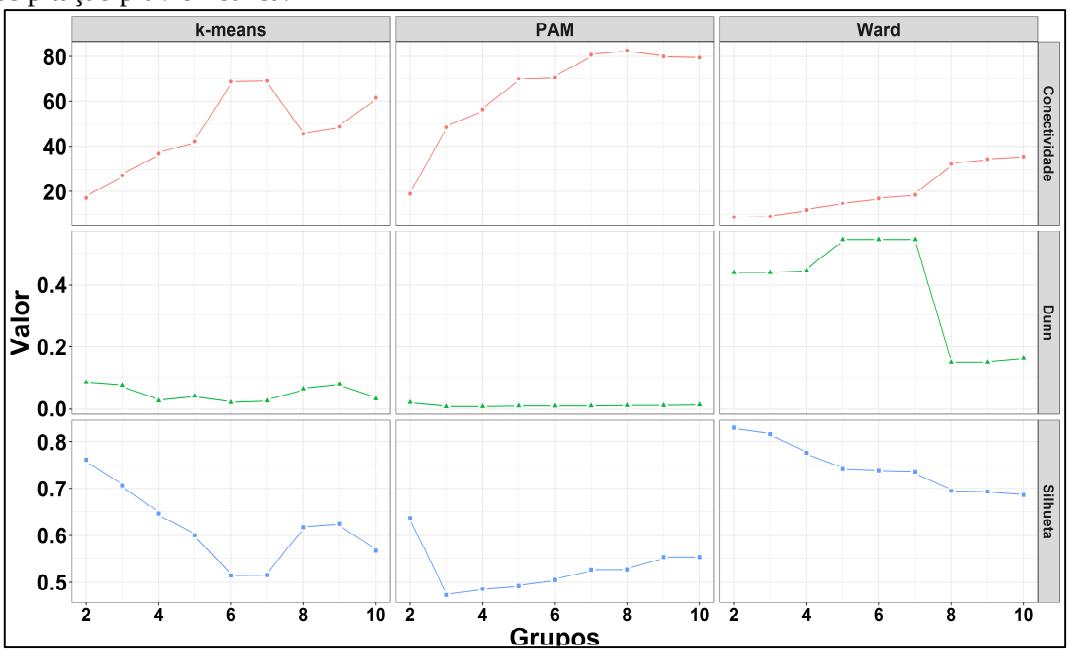
ZHANG, S. et al. Learning k for kNN Classification. **ACM Transactions on Intelligent Systems and Technology**, v. 8, n. 3, p. 1–19, 12 jan. 2017.

ZHAO, W.; BEACH, T. H.; REZGUI, Y. Automated Model Construction for Combined Sewer Overflow Prediction Based on Efficient LASSO Algorithm. **IEEE Transactions on Systems, Man, and Cybernetics: Systems**, p. 1–16, 2017.

APÊNDICES - A

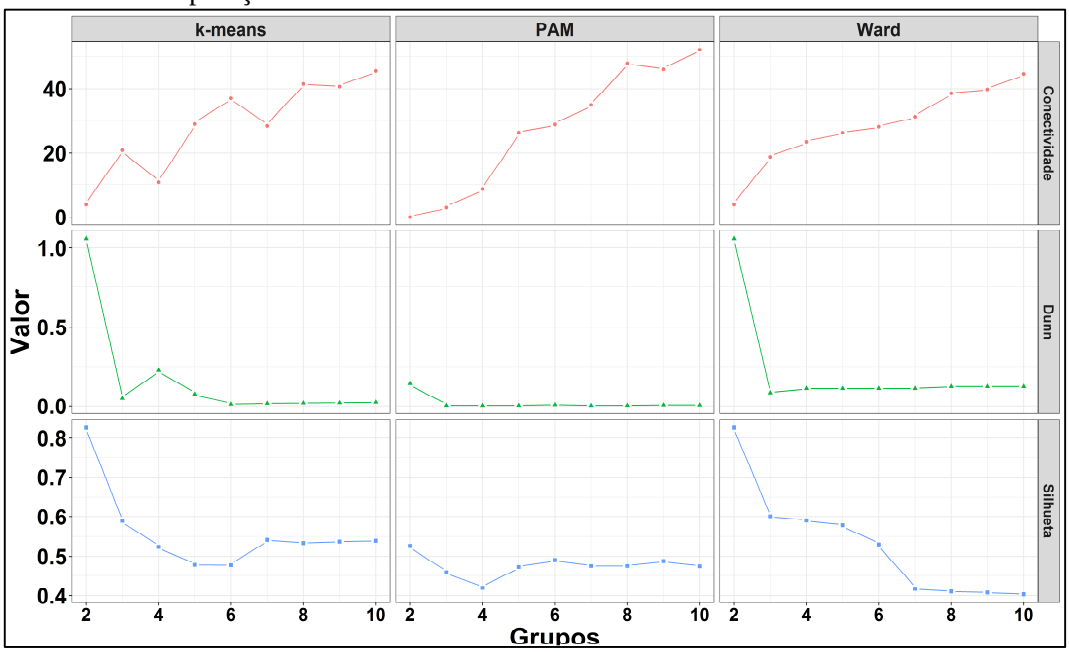
Avaliação para validação interna de agrupamentos considerando as métricas de Conectividade, índice Dunn e Silhueta.

Figura A.1 – Apêndices – Validação interna para grupos de objetos relacionados a precipitação pluviométrica.



Fonte: O Autor, 2019.

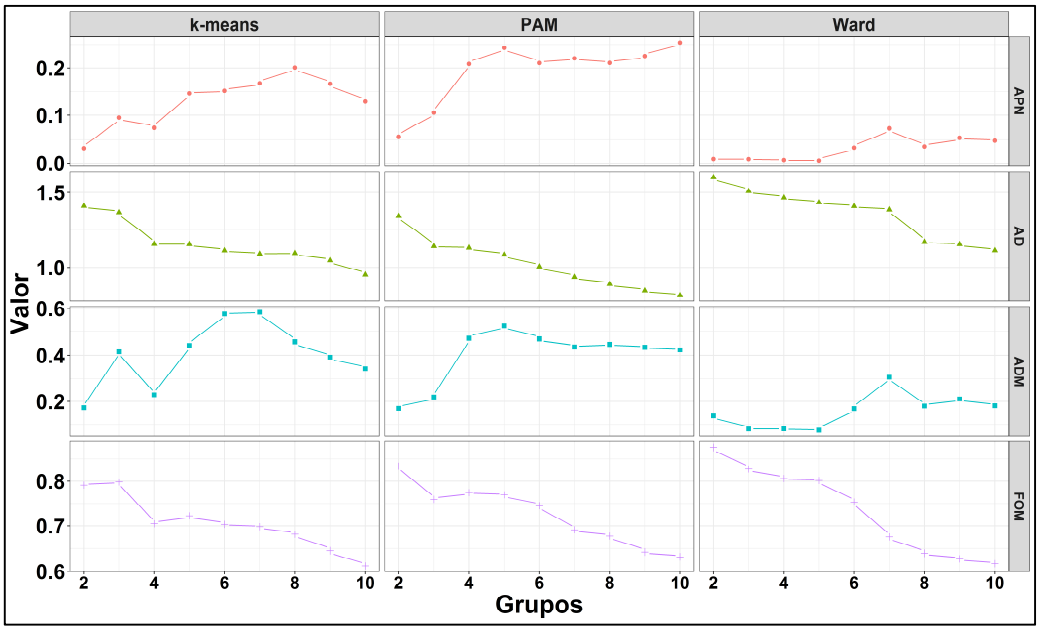
Figura A.2 – Apêndices – Validação interna para grupos de objetos relacionados ao transbordamento/operação.



Fonte: O Autor, 2019.

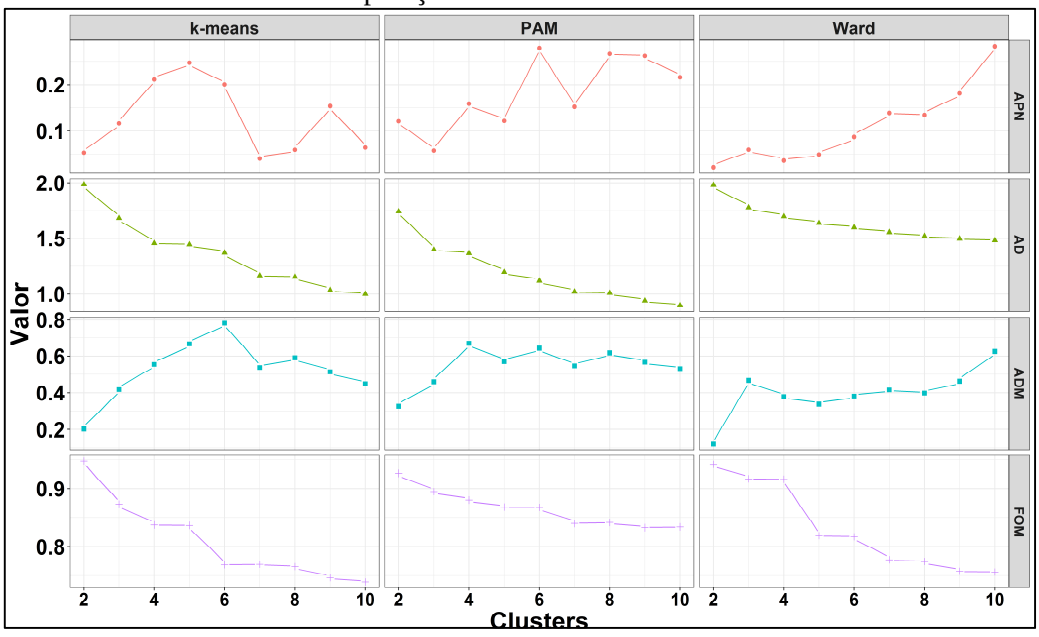
Avaliação para validação de estabilidade de agrupamentos, considerando as métricas de proporção média de observações não classificadas no mesmo grupos (*average porportion of nonoverlap* - APN), distância média entre objetos em um mesmo grupo (*average distance* - AD), distância média entre os centroides quando as observações estão no mesmo grupo (*average distance between means* - ADM) e figura de mérito (*figure of merit* - FOM).

Figura A.3 – Apêndices – Avaliação de estabilidade para grupos de objetos relacionados a chuva.



Fonte: O Autor, 2019.

Figura A.4 – Apêndices – Avaliação de estabilidade para grupos de objetos relacionados ao transbordamento/operação.



Fonte: O Autor, 2019.

Tabelas com resultados das análises de validação interna e estabilidade de agrupamentos.

Tabela A.1 - Validação interna para grupos de dados relacionados a precipitação pluviométrica.

Algoritmo	Métrica	Número de Grupos								
		2	3	4	5	6	7	8	9	10
Hierárquico	Conectividade	8,7698	8,9948	11,8238	14,7528	17,0028	18,5028	32,3841	34,3841	35,4492
Hierárquico	Dunn	0,4388	0,4388	0,4439	0,5463	0,5463	0,5463	0,1486	0,1486	0,1620
Hierárquico	Silhueta	0,8298	0,8168	0,7760	0,7426	0,7391	0,7366	0,6946	0,6930	0,6858
k-means	Conectividade	17,2448	27,2980	36,8667	42,1012	68,5159	68,8492	45,4246	48,8579	61,5349
k-means	Dunn	0,0830	0,0724	0,0254	0,0400	0,0201	0,0250	0,0618	0,0769	0,0314
k-means	Silhueta	0,7617	0,7065	0,6466	0,6007	0,5137	0,5145	0,6175	0,6246	0,5663
PAM	Conectividade	18,8984	48,5659	56,2361	69,8500	70,5698	80,6885	82,5210	79,9857	79,5194
PAM	Dunn	0,0208	0,0074	0,0077	0,0097	0,0096	0,0097	0,0109	0,0109	0,0133
PAM	Silhueta	0,6369	0,4721	0,4854	0,4926	0,5047	0,5258	0,5262	0,5526	0,5525

Fonte: O Autor, 2019.

Tabela A.2 - Validação interna para grupos de dados relacionados ao transbordamento/operação.

Algoritmo	Métrica	Número de Grupos								
		2	3	4	5	6	7	8	9	10
Hierárquico	Conectividade	3,8579	18,7520	23,4111	26,2052	28,0718	31,1008	38,6440	39,8107	44,5897
Hierárquico	Dunn	1,0571	0,0862	0,1129	0,1129	0,1129	0,1129	0,1253	0,1253	0,1253
Hierárquico	Silhueta	0,8259	0,6006	0,5899	0,5779	0,5299	0,4169	0,4101	0,4076	0,4033
k-means	Conectividade	3,8579	20,8770	10,7262	28,9016	37,0246	28,2917	41,5587	40,7210	45,5000
k-means	Dunn	1,0571	0,0506	0,2257	0,0786	0,0126	0,0179	0,0200	0,0220	0,0252
k-means	Silhueta	0,8259	0,5888	0,5239	0,4784	0,4776	0,5416	0,5333	0,5370	0,5392
PAM	Conectividade	0,0000	2,9063	8,6123	26,2468	28,7385	34,8075	47,8881	45,9845	52,1032
PAM	Dunn	0,1441	0,0046	0,0038	0,0043	0,0092	0,0034	0,0041	0,0075	0,0069
PAM	Silhueta	0,5262	0,4585	0,4192	0,4726	0,4906	0,4753	0,4753	0,4882	0,4746

Fonte: O Autor, 2019.

Tabela A.3 - Validação da estabilidade entre grupos de dados relacionados a precipitação pluviométrica.

Algoritmo	Métrica	Número de Grupos								
		2	3	4	5	6	7	8	9	10
Hierárquico	APN	0,0089	0,0087	0,0068	0,0055	0,0331	0,0725	0,0353	0,0536	0,0479
Hierárquico	AD	1,6015	1,5085	1,4652	1,4324	1,4069	1,3838	1,1712	1,1566	1,1156
Hierárquico	ADM	0,1369	0,0805	0,0810	0,0753	0,1690	0,3054	0,1801	0,2105	0,1814
Hierárquico	FOM	0,8746	0,8276	0,8061	0,8018	0,7534	0,6753	0,6398	0,6280	0,6175
k-means	APN	0,0318	0,0947	0,0736	0,1465	0,1518	0,1677	0,2011	0,1665	0,1288
k-means	AD	1,4083	1,3650	1,1587	1,1571	1,1149	1,0921	1,0967	1,0475	0,9517
k-means	ADM	0,1730	0,4157	0,2284	0,4416	0,5763	0,5836	0,4570	0,3921	0,3423
k-means	FOM	0,7912	0,7983	0,7047	0,7224	0,7034	0,6980	0,6813	0,6452	0,6124
PAM	APN	0,0556	0,1052	0,2089	0,2432	0,2110	0,2207	0,2111	0,2254	0,2542
PAM	AD	1,3403	1,1435	1,1333	1,0860	1,0059	0,9358	0,8878	0,8446	0,8120
PAM	ADM	0,1701	0,2186	0,4716	0,5240	0,4690	0,4358	0,4465	0,4344	0,4238
PAM	FOM	0,8340	0,7594	0,7745	0,7696	0,7451	0,6911	0,6771	0,6425	0,6323

Fonte: O Autor, 2019.

Tabela A.4 - Validação da estabilidade entre grupos de dados relacionados ao transbordamento/operação.

Algoritmo	Métrica	Número de Grupos								
		2	3	4	5	6	7	8	9	10
Hierárquico	APN	0,0222	0,0598	0,0361	0,0490	0,0871	0,1388	0,1344	0,1811	0,2830
Hierárquico	AD	1,9802	1,7791	1,6953	1,6345	1,5953	1,5530	1,5191	1,4968	1,4821
Hierárquico	ADM	0,1211	0,4676	0,3781	0,3378	0,3808	0,4160	0,3978	0,4631	0,6248
Hierárquico	FOM	0,9417	0,9172	0,9164	0,8193	0,8174	0,7778	0,7748	0,7576	0,7561
k-means	APN	0,0530	0,1165	0,2120	0,2473	0,2007	0,0404	0,0597	0,1531	0,0645
k-means	AD	1,9850	1,6796	1,4557	1,4439	1,3676	1,1634	1,1513	1,0341	1,0005
k-means	ADM	0,2022	0,4179	0,5548	0,6654	0,7823	0,5369	0,5913	0,5132	0,4489
k-means	FOM	0,9478	0,8733	0,8375	0,8367	0,7691	0,7700	0,7659	0,7461	0,7386
PAM	APN	0,1212	0,0579	0,1582	0,1220	0,2794	0,1516	0,2670	0,2623	0,2166
PAM	AD	1,7424	1,3953	1,3636	1,1961	1,1157	1,0211	1,0094	0,9390	0,8952
PAM	ADM	0,3266	0,4600	0,6683	0,5705	0,6421	0,5462	0,6172	0,5673	0,5305
PAM	FOM	0,9264	0,8947	0,8804	0,8685	0,8681	0,8401	0,8418	0,8330	0,8343

Fonte: O Autor, 2019.

Matrizes de confusão referente aos modelos preditivos criados.

Tabela A.5 - Matriz de confusão para KNN sem técnica de amostragem no cenário 1.

Evento Predito	Evento Observado	
	Extravasamento	Operação Normal
Extravasamento	4	3
Operação Normal	5	51

Fonte: O Autor, 2019.

Tabela A.6 - Matriz de confusão para KNN com *undersampling* no cenário 1.

Evento Predito	Evento Observado	
	Extravasamento	Operação Normal
Extravasamento	9	8
Operação Normal	0	46

Fonte: O Autor, 2019.

Tabela A.7 - Matriz de confusão para KNN com *oversampling* no cenário 1.

Evento Predito	Evento Observado	
	Extravasamento	Operação Normal
Extravasamento	6	4
Operação Normal	3	50

Fonte: O Autor, 2019.

Tabela A.8 - Matriz de confusão para KNN com ROSE no cenário 1.

Evento Predito	Evento Observado	
	Extravasamento	Operação Normal
Extravasamento	9	5
Operação Normal	0	49

Fonte: O Autor, 2019.

Tabela A.9 - Matriz de confusão para *Random Forest* sem técnica de amostragem no cenário 1.

Evento Predito	Evento Observado	
	Extravasamento	Operação Normal
Extravasamento	7	2
Operação Normal	2	52

Fonte: O Autor, 2019.

Tabela A.10 - Matriz de confusão para *Random Forest* com *undersampling* no cenário 1.

Evento Predito	Evento Observado	
	Extravasamento	Operação Normal
Extravasamento	9	11
Operação Normal	0	43

Fonte: O Autor, 2019.

Tabela A.11 - Matriz de confusão para *Random Forest* com *oversampling* no cenário 1.

Evento Predito	Evento Observado	
	Extravasamento	Operação Normal
Extravasamento	8	3
Operação Normal	1	51

Fonte: O Autor, 2019.

Tabela A.12 - Matriz de confusão para *Random Forest* com ROSE no cenário 1.

Evento Predito	Evento Observado	
	Extravasamento	Operação Normal
Extravasamento	9	8
Operação Normal	0	46

Fonte: O Autor, 2019.

Tabela A.13 - Matriz de confusão para KNN sem técnica de amostragem no cenário 2.

Evento Predito	Evento Observado	
	Extravasamento	Operação Normal
Extravasamento	3	1
Operação Normal	6	53

Fonte: O Autor, 2019.

Tabela A.14 - Matriz de confusão para KNN com *undersampling* no cenário 2.

Evento Predito	Evento Observado	
	Extravasamento	Operação Normal
Extravasamento	7	5
Operação Normal	2	49

Fonte: O Autor, 2019.

Tabela A.15 - Matriz de confusão para KNN com *oversampling* no cenário 2.

Evento Predito	Evento Observado	
	Extravasamento	Operação Normal
Extravasamento	5	2
Operação Normal	4	52

Fonte: O Autor, 2019.

Tabela A.16 - Matriz de confusão para KNN com ROSE no cenário 2.

Evento Predito	Evento Observado	
	Extravasamento	Operação Normal
Extravasamento	4	0
Operação Normal	5	54

Fonte: O Autor, 2019.

Tabela A.17 - Matriz de confusão para *Random Forest* sem técnica de amostragem no cenário 2.

Evento Predito	Evento Observado	
	Extravasamento	Operação Normal
Extravasamento	5	2
Operação Normal	4	52

Fonte: O Autor, 2019.

Tabela A.18 - Matriz de confusão para *Random Forest* com *undersampling* no cenário 2.

Evento Predito	Evento Observado	
	Extravasamento	Operação Normal
Extravasamento	7	6
Operação Normal	2	48

Fonte: O Autor, 2019.

Tabela A.19 - Matriz de confusão para *Random Forest* com *oversampling* no cenário 2.

Evento Predito	Evento Observado	
	Extravasamento	Operação Normal
Extravasamento	6	4
Operação Normal	3	50

Fonte: O Autor, 2019.

Tabela A.20 - Matriz de confusão para *Random Forest* com ROSE no cenário 2.

Evento Predito	Evento Observado	
	Extravasamento	Operação Normal
Extravasamento	9	9
Operação Normal	0	45

Fonte: O Autor, 2019.

Tabela A.21 - Matriz de confusão para KNN sem técnica de amostragem no cenário 3.

Evento Predito	Evento Observado	
	Extravasamento	Operação Normal
Extravasamento	2	0
Operação Normal	7	54

Fonte: O Autor, 2019.

Tabela A.22 - Matriz de confusão para KNN com *undersampling* no cenário 3.

Evento Predito	Evento Observado	
	Extravasamento	Operação Normal
Extravasamento	9	5
Operação Normal	0	49

Fonte: O Autor, 2019.

Tabela A.23 - Matriz de confusão para KNN com *oversampling* no cenário 3.

Evento Predito	Evento Observado	
	Extravasamento	Operação Normal
Extravasamento	4	0
Operação Normal	5	54

Fonte: O Autor, 2019.

Tabela A.24 - Matriz de confusão para KNN com ROSE no cenário 3.

Evento Predito	Evento Observado	
	Extravasamento	Operação Normal
Extravasamento	3	1
Operação Normal	6	53

Fonte: O Autor, 2019.

Tabela A.25 - Matriz de confusão para *Random Forest* sem técnica de amostragem no cenário 3.

Evento Predito	Evento Observado	
	Extravasamento	Operação Normal
Extravasamento	5	1
Operação Normal	4	53

Fonte: O Autor, 2019.

Tabela A.26 - Matriz de confusão para *Random Forest* com *undersampling* no cenário 3.

Evento Predito	Evento Observado	
	Extravasamento	Operação Normal
Extravasamento	9	6
Operação Normal	0	48

Fonte: O Autor, 2019.

Tabela A.27- Matriz de confusão para *Random Forest* com *oversampling* no cenário 3.

Evento Predito	Evento Observado	
	Extravasamento	Operação Normal
Extravasamento	5	2
Operação Normal	4	52

Fonte: O Autor, 2019.

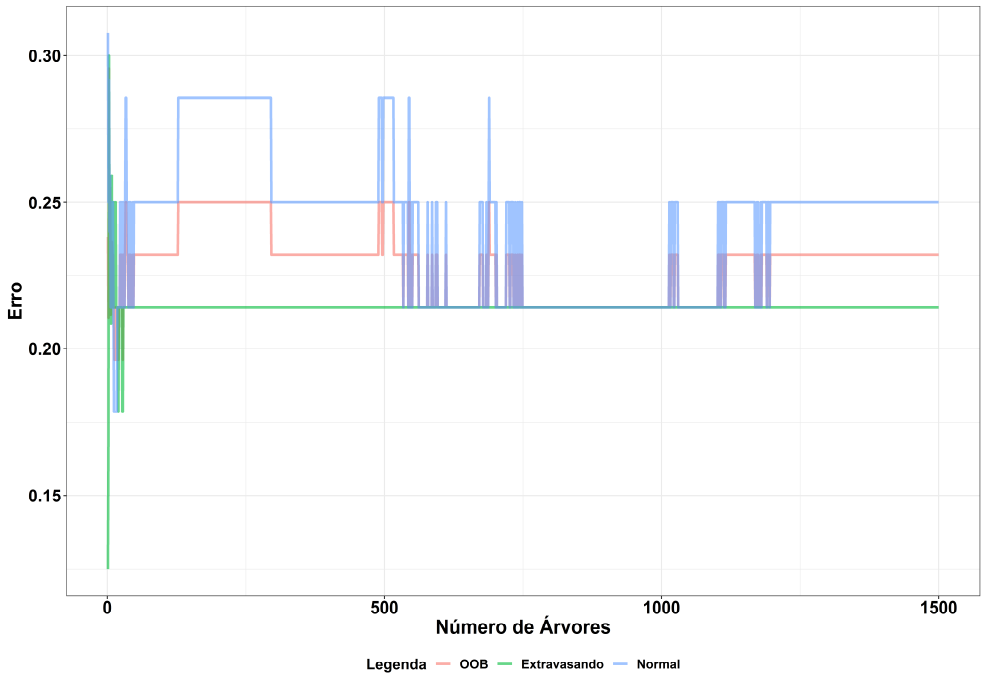
Tabela A.28 - Matriz de confusão para *Random Forest* com ROSE no cenário 3.

Evento Predito	Evento Observado	
	Extravasamento	Operação Normal
Extravasamento	7	2
Operação Normal	2	52

Fonte: O Autor, 2019.

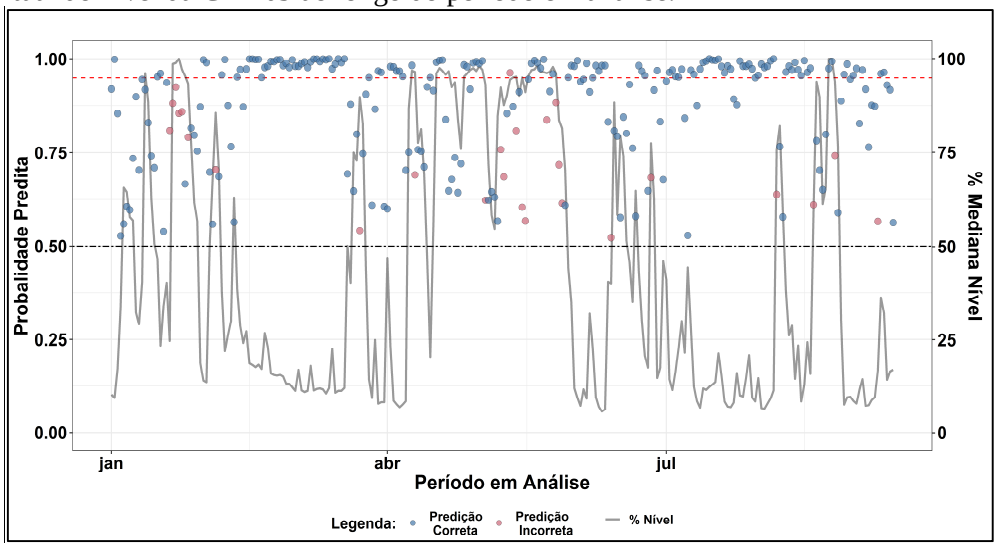
Out-of-bag error (OOB), Figura 5 – Apêndice, e probabilidade de extravasamento ao longo do período em análise, Figura 6 – Apêndices, indicadas a partir do modelo *Random Forest* com *undersampling* no cenário 3.

Figura A.5 – Apêndices – OOB do modelo de predição em escolhido (*Random Forest* com *undersampling*).



Fonte: O Autor, 2019.

Figura A.6 – Apêndices – Probabilidade de extravasamento, classificação da predição e percentual do nível da CET 09 ao longo do período em análise.



Fonte: O Autor, 2019.

UFBA
UNIVERSIDADE FEDERAL DA BAHIA
ESCOLA POLITÉCNICA

PROGRAMA DE PÓS- GRADUAÇÃO EM ENGENHARIA INDUSTRIAL - PEI

Rua Aristides Novis, 02, 6º andar, Federação, Salvador BA
CEP: 40.210-630
Telefone: (71) 3283-9800
E-mail: pei@ufba.br
Home page: http://www.pei.ufba.br