



Serviço Público Federal
Ministério da Educação
Fundação Universidade Federal de Mato Grosso do Sul



FUNDAÇÃO UNIVERSIDADE FEDERAL DE MATO GROSSO DO SUL
FACULDADE DE ENGENHARIAS, ARQUITETURA E URBANISMO E GEOGRAFIA
MESTRADO PROFISSIONAL EM EFICIÊNCIA ENERGÉTICA E SUSTENTABILIDADE

DETECÇÃO DE FRAUDES EM SISTEMAS DE
ABASTECIMENTO DE ÁGUA POR MÁQUINAS DE VETORES
DE SUPORTE - SVM

ELTHON SANTOS TEIXEIRA

CAMPO GRANDE - MS

2018

**FUNDAÇÃO UNIVERSIDADE FEDERAL DE MATO GROSSO DO SUL
FACULDADE DE ENGENHARIAS, ARQUITETURA E URBANISMO E GEOGRAFIA
MESTRADO PROFISSIONAL EM EFICIÊNCIA ENERGÉTICA E SUSTENTABILIDADE**

**DETECÇÃO DE FRAUDES EM SISTEMAS DE
ABASTECIMENTO DE ÁGUA POR MÁQUINAS DE VETORES
DE SUPORTE - SVM**

ELTHON SANTOS TEIXEIRA

Trabalho apresentado para a etapa de defesa do Mestrado Profissional na Faculdade de Engenharias, Arquitetura e Urbanismo e Geografia da Universidade Federal de Mato Grosso do Sul.

Orientador: Prof. Dr. Peter Cheung Batista

**CAMPO GRANDE - MS
JANEIRO/ 2018**

ELTHON SANTOS TEIXEIRA

DETECÇÃO DE FRAUDES EM SISTEMAS DE ABASTECIMENTO DE ÁGUA POR MÁQUINAS DE VETORES DE SUPORTE - SVM

RESUMO

TEIXEIRA, E. S., 2017.

DETECÇÃO DE FRAUDES EM SISTEMAS DE ABASTECIMENTO DE ÁGUA POR MÁQUINAS DE VETORES DE SUPORTE - SVM

Diversos são os fatores relacionados às causas das perdas de água nos Sistemas de Abastecimento de Água e, conseqüentemente, à redução de sua eficiência energética. Em 2014, o segundo item das despesas com explorações das empresas de Saneamento, ficou para as despesas com energia elétrica, tendo totalizado naquele ano o valor de R\$ 3.471,0 bilhões, ou seja 11,2% do total. A relevância nesse processo se dá pela necessidade de mover água contra a ação da gravidade em tubulações pressurizadas utilizando conjuntos moto bomba. Portanto, a redução desses custos é um objetivo a ser perseguido, partindo do planejamento prévio para implantação até às etapas rotineiras de operação e manutenção. Um fator impactante sobre os custos de energia para as companhias de saneamento é a perda em um sistema de abastecimento. Nos sistemas públicos de abastecimento as perdas de água correspondem aos volumes não contabilizados, que podem ser físicas ou reais, que representam a parcela não consumida, quanto as perdas não físicas ou aparentes, que correspondem à água consumida e não contabilizada. Especificamente sobre as perdas aparentes, uma das principais causas se dá pelas fraudes, ou consumo não autorizado de água tratada pela população. Uma forma de reduzir tais perdas, é aumentar a assertividade das companhias nas ações de pesquisa destas ligações irregulares, que atualmente são feitas praticamente ao acaso, utilizando métodos lentos de detecção. Para isso é proposta uma análise de séries de dados temporais de consumo e correlação com históricos de dados obtidos no Sistema de Informações comerciais de uma Companhia de Saneamento do Brasil, para desenvolver um modelo de previsão e testar a sua margem de erro, baseado em métodos de Mineração de Dados, Outliers e Clusterização. A hipótese desenvolvida neste estudo é de que é possível traçar, através da metodologia de inteligência artificial, especificamente máquinas de vetores de suporte – SVM, através de séries temporais, ou seja, dados históricos de consumo das ligações de uma determinada localidade, um perfil que seja comum a parcela de consumidores que praticam o consumo não autorizado. Através da análise destes dados históricos, podem ser criados mecanismos automatizados de identificação de ligações fraudadas, e assim combater esta prática, reduzindo o consumo não autorizado e os custos de energia vinculados a perdas de água tratada.

Palavras-chaves: Mineração de Dados, Perdas Aparentes, Água Tratada, Detecção de Fraudes, Inteligência Artificial, Máquinas de Vetores de Suporte.

ABSTRACT

There are several factors related to the causes of water losses in Water Supply Systems and, consequently, the reduction of their energy efficiency. In 2014, the second item of expenses with explorations of the Sanitation utilities, was for the costs with electric energy, having totaled in that year the value of R\$ 3,471.0 billion, that is 11.2% of the total. The relevance in this process is due to the need to move water against the action of gravity in pressurized pipes using motor pump sets. Therefore, the reduction of these costs is an objective to be pursued, starting from the previous implementation planning and the routine steps of operation and maintenance. An impacting factor on energy costs for sanitation companies are water the losses in a supply system. In public water supply systems, water losses correspond to unaccounted volumes, which can be physical or real, representing the unconsumed portion, as well as non-physical or apparent losses, which correspond to consumed and unaccounted water. Specifically, regarding apparent losses, one of the main causes is due to fraud, or unauthorized consumption of treated water by the population. One way to reduce such losses is to increase the companies' assertiveness in researching these irregular connections, which are currently done practically at random, using slow detection methods. For this purpose, an analysis of temporal consumption data series and correlation with historical data obtained from the Commercial Information System of a Sanitation Company in Brazil is proposed, to develop a forecast model and test its margin of error, based on methods of Data Mining, Outliers and Clustering. The hypothesis developed in this study is that it is possible to trace, through the methodology of artificial intelligence, specifically support vector machines - SVM, through time series, that are, historical data of consumption of the connections of a certain location, a profile that is common to the share of consumers who practice unauthorized consumption. Through the analysis of this historical data, automated mechanisms can be created to identify fraudulent connections, and thus combat this practice, reducing unauthorized consumption and energy costs linked to the loss of treated water.

Keywords: Data Mining, Apparent Losses, Treated Water, Fraud Detection, Artificial Intelligence, Support Vector Machines.

LISTA DE ILUSTRAÇÕES

Figura 01 - Composição Média das despesas de exploração – DEX (Informação FN015) dos prestadores de serviços participantes do SNIS em 2014.	8
Figura 02 – Principais componentes de Despesa de Exploração – DEX (Informação FN015) dos prestadores de serviços participantes do SNIS em Mato Grosso do Sul nos últimos cinco anos.....	9
Figura 03 – Distribuição espacial das fraudes no DMC 056.....	34
Figura 04 – Distribuição espacial das fraudes no DMC 057.....	35
Figura 05 – Distribuição espacial das fraudes no DMC 059.....	35
Figura 06 – Distribuição espacial das fraudes no DMC 060.....	36
Figura 07 – Distribuição das médias de consumo por classe (DMCs 056, 057, 059 e 060).....	37
Figura 08 – Ocorrências comerciais x Fraudes x Consumo zero – DMC-056.....	38
Figura 09 – Ocorrências comerciais x Fraudes x Consumo zero – DMC-057.....	39
Figura 10 – Ocorrências comerciais x Fraudes x Consumo zero – DMC-059.....	39
Figura 11 – Ocorrências comerciais x Fraudes x Consumo zero – DMC-060.....	40
Figura 12 – Precisão obtida através dos grupos de dados propostos.....	41
Figura 13 – Precisão obtida sem o uso de subamostragem.	41
Figura 14 – Precisão obtida com o uso de subamostragem (Sub-10 e Sub-50).	42
Figura 15 – Precisão obtida com o uso de subamostragem (Sub-10 e Sub-5).	42
Figura 16 – Precisão obtida com o uso de subamostragem (Sub-5 e Sub-50/50).	43
Figura 17 – Revocação obtida através dos grupos de dados propostos.....	44
Figura 18 – Revocação obtida sem o uso de subamostragem.	44
Figura 19 – Revocação obtida com o uso de subamostragem (Sub-10 e Sub-50).	45
Figura 20 – Revocação obtida com o uso de subamostragem (Sub-10 e Sub-5).	45
Figura 21 – Revocação obtida com o uso de subamostragem (Sub-5 e Sub-50/50).	46

LISTA DE TABELAS:

Tabela 01 – Dados das áreas de estudo.....	30
Tabela 02 – Resultados de precisão para os grupos de dados completos e após subamostragem.	41
Tabela 03 – Resultados de revocação para os grupos de dados completos e após subamostragem.	43

SUMÁRIO

1	INTRODUÇÃO	6
1.1	ASPECTOS INTRODUTÓRIOS DO TEMA	9
1.1.1	Tema.....	9
1.1.2	Contextualização da problemática	9
1.1.3	Problema central da pesquisa	10
1.1.4	Hipóteses da pesquisa	11
1.1.5	Justificativas da pesquisa	11
1.2	OBJETIVOS	11
1.2.1	Objetivo Geral.....	11
1.2.2	Objetivos Específicos.....	12
2	REVISÃO BIBLIOGRÁFICA	13
2.1	Eficiência Energética em Sistemas de Saneamento.....	13
2.2	Redução de Custos com Energia em Saneamento	14
2.3	Perdas de Água em Sistemas de Abastecimento Público	16
2.3.1	Perdas reais de água	17
2.3.2	Perdas de água aparentes.....	19
2.4	Mineração de Dados e a Identificação de Perdas Aparentes	23
2.4.1	Algoritmos de Classificação	27
3	METODOLOGIA.....	31
3.1	Delineamento da pesquisa.....	31
3.2	Aquisição de dados	32
3.3	Seleção e extração dos dados.....	32
3.4	Normalização de dados	34
3.5	SVM Classificador e otimização	34
4	RESULTADOS E DISCUSSÕES	37

4.1	Características	37
4.1.1	Geolocalização	37
4.1.2	Médias de Consumo por Classes	39
4.1.3	Fraudes por Ocorrências e Consumo Zero.....	40
4.2	Verificação de Desempenho do Algoritmo	43
4.2.1	Precisão	43
4.2.2	Revocação	46
5	CONCLUSÃO.....	50
	REFERÊNCIAS	52

1 INTRODUÇÃO

De acordo com a Empresa de Pesquisa Energética (EPE), órgão ligado ao Ministério de Minas e Energia (MMA), a capacidade instalada de geração de energia no mundo em 2014 era de 6.038,7 GW. O mesmo órgão aponta que as Américas do Sul e Central são responsáveis por 5,2% desse montante, com 313,8 GW (BRASIL, 2017).

Ainda segundo a EPE, em seu Anuário Estatístico de Energia Elétrica – 2017 (BRASIL, 2017), nesse panorama, o Brasil se sobressai comparado aos demais países da América Latina, ocupando em 2014 a oitava posição no ranking dos dez maiores países, com uma capacidade instalada de geração de 133,9 GW, ou seja, 2,2% da capacidade mundial.

Do ponto de vista de energia renovável, esta posição de destaque se reafirma considerando que o Brasil em 2014 foi o terceiro maior produtor de energia hidrelétrica do mundo, estando apenas atrás da China e Canadá, com 373,4 TWh, representando 9,8% do total produzido pelos dez maiores países.

Infelizmente, esta abundância em energia não se reflete nos valores das tarifas em relação aos demais países. O próprio EPE (BRASIL, 2014) apontava que o Brasil possuía a terceira tarifa mais cara do mundo, atrás apenas da Itália e Japão. O valor médio da tarifa industrial em 2012 foi de U\$ 177,68/MWh, ou R\$ 257,30 /MWh. Esse valor era três vezes maior que a tarifa de países desenvolvidos, como EUA e Noruega.

No ano de 2014, foi aprovado pela ANEEL a redução das tarifas de energia como resultado da Lei nº 12.783/2013, que promoveu a renovação das concessões de transmissão e geração de energia que venciam até 2017, e das medidas provisórias 591/2012 e 605/2013. Para os consumidores de alta tensão, o desconto chegaria a 32%. Esta ação do Governo Federal manteve a tarifa média industrial em R\$ 249,01/MWh no ano de 2014 porém, em 2015 este valor subiu para R\$ 335,31/MWh, e novamente em 2016 voltou a subir para R\$ 392,83/MWh, ou seja, um aumento aproximado de 157% em um período de dois anos (BRASIL, 2017).

O setor industrial do ponto de vista do consumo é o grande consumidor de energia no país e em 2016 representou 35,7% dos 460.829 GWh consumidos (BRASIL, 2017).

Observa-se, entretanto, que na região Centro-Oeste, que representa 7,5% do consumo de energia no país, essa realidade é um tanto diferente. Por não possuir um nível de industrialização como as regiões sudeste e sul, por exemplo, o seu maior consumo é residencial, mas ainda assim seguido de perto pela atividade industrial (BRASIL, 2017).

No Mato Grosso do Sul, de acordo com o Portal da Indústria (CNI, 2014), a indústria do Estado paga a 8ª tarifa de energia mais elevada entre as unidades da Federação, sendo 4,4% maior que a média nacional. A classe Industrial no Mato Grosso do Sul é responsável por 12,5% do consumo industrial na região Centro-Oeste, com 1.083 GWh consumidos por 8.742 consumidores industriais em 2016 (BRASIL, 2017).

Cabe, portanto, aos empreendedores do setor industrial buscar alternativas para reduzir o consumo e os custos com energia. Sendo esta consumida de forma ineficiente, vai de encontro a necessidade de sustentabilidade em relação aos recursos naturais, e que contratados de forma equivocada, sob o ponto de vista econômico, a curto prazo oneram o valor final de seus produtos aos consumidores e aumentam a carga tributária paga pelas empresas, respectivamente. Quanto ao aspecto ambiental, em longo prazo pode inclusive inviabilizar o processo produtivo com a escassez dos recursos gerada pelo seu mau uso.

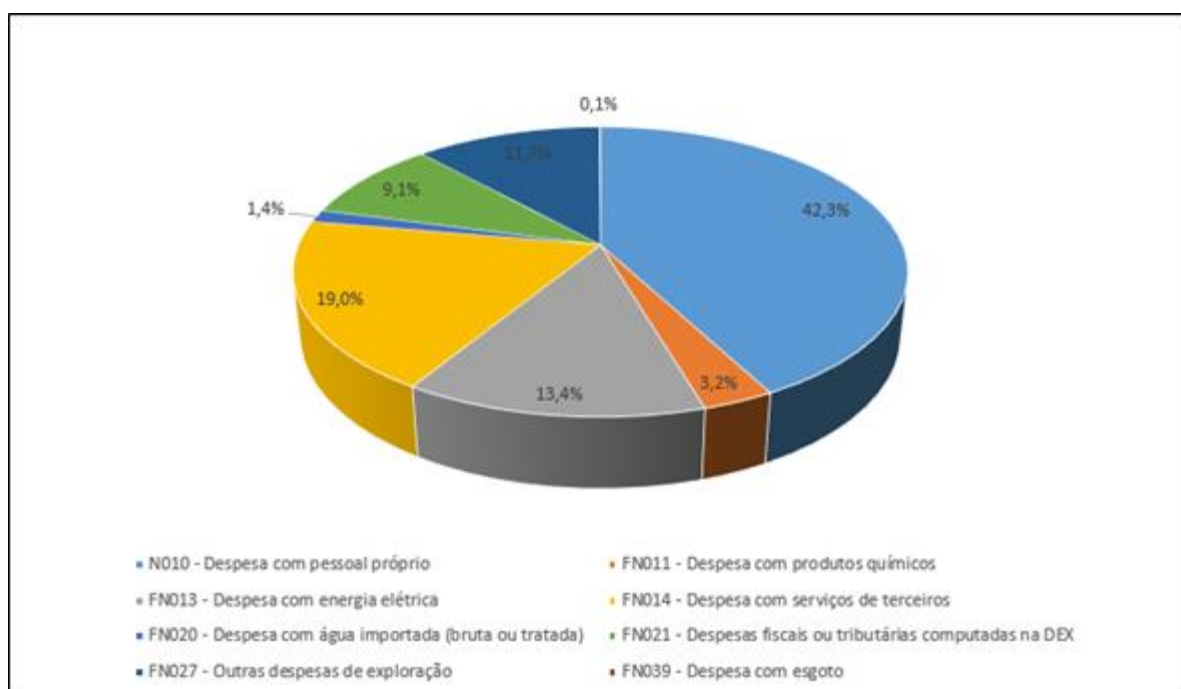


Figura 1 - Composição Média das despesas de exploração – DEX (Informação FN015) dos prestadores de serviços participantes do SNIS em 2014. (SNIS,2015)

Nesse aspecto se enquadra o tema deste projeto. Dentre os grandes setores do segmento industrial no Brasil, destaca-se pela importância social, econômica e ambiental o setor do Saneamento. Sua relevância para os estudos na área de pesquisa do Programa de Eficiência Energética e Sustentabilidade de Mato Grosso do Sul se observa, principalmente pela sua

relação com o uso de energia, que permeia assuntos ambientais, econômicos e sociais, ou seja, sustentáveis. Tal fato é confirmado por David Zee, quando o mesmo afirma que “*a ciência já tem conhecimento da importância do saneamento básico no contexto da sustentabilidade urbana*” (TRATABRASIL, 2014).

De acordo com o Sistema Nacional de Informações do Saneamento (SNIS), em 2014, o segundo item das Despesas com Explorações das Empresas de Saneamento (DEX), ficou para as despesas com energia elétrica (considerando serviços de terceiros e pessoal como insumo mão de obra), tendo totalizado naquele ano o valor de R\$ 37,8 bilhões, ou seja 13,4% do total, conforme observado na figura 01 (BRASIL, 2016).

Essa relevância no processo se dá pelo fato de que, no setor do saneamento, de acordo com Andrade Filho (2009), “é comum a necessidade de mover água e ou esgoto contra a ação da gravidade em tubulações pressurizadas”.

No Mato Grosso do Sul, o cenário não é diferente do observado no âmbito nacional. De acordo com o SNIS (BRASIL, 2015), o terceiro item das Despesas de Exploração das empresas de Saneamento que atuam no abastecimento de água tratada e coleta e tratamento de esgoto sanitário, atrás da mão de obra e despesas fiscais e tributárias, é a despesa com energia elétrica, tendo totalizado em 2014 o valor de R\$ 50.235.116,38, ou seja 11,06% das despesas de exploração, como observa-se na figura 2.

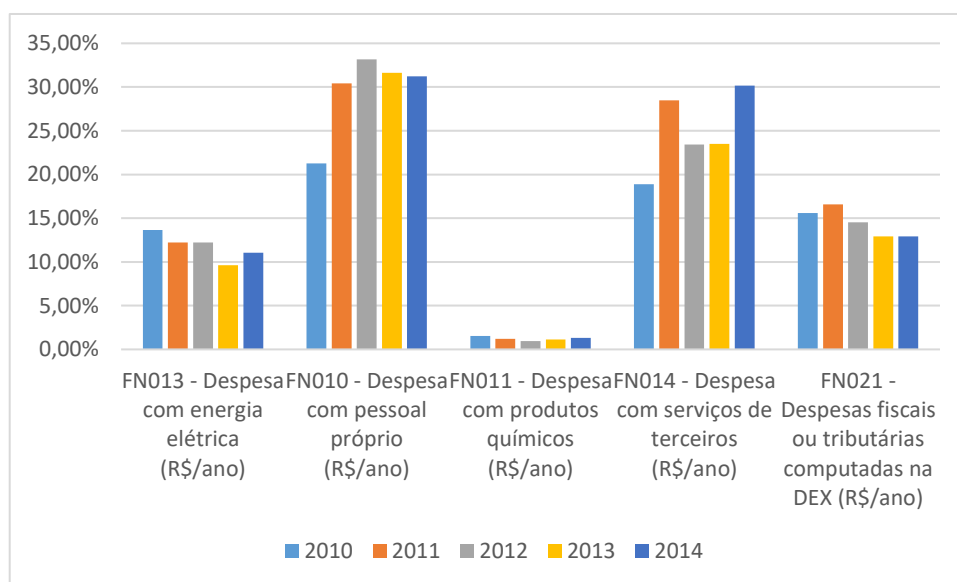


Figura 2 – Principais componentes de Despesa de Exploração – DEX (Informação FN015) dos prestadores de serviços participantes do SNIS em Mato Grosso do Sul nos últimos cinco anos. (SNIS,2015)

Esses custos são diretamente relacionados as perdas de água em sistemas de abastecimento, pois os volumes não contabilizados, perdidos na distribuição, entre a produção e o consumo, demanda energia (sejam vazamentos ou consumos não autorizados) para serem bombeados.

1.1 ASPECTOS INTRODUTÓRIOS DO TEMA

1.1.1 Tema

O tema delimitado foi a “DETECÇÃO DE FRAUDES EM SISTEMAS DE ABASTECIMENTO DE ÁGUA POR MÁQUINAS DE VETORES DE SUPORTE - SVM”.

1.1.2 Contextualização da problemática

Jansen e Schulz (2006), apresentam o problema da escassez de água no mundo, com foco em países em desenvolvimento como é o caso do Brasil. Os autores afirmam que a água é importante para o desenvolvimento na medida em que o acesso a este recurso, água limpa e saneamento, melhora a saúde e a produtividade das comunidades. O atendimento a esta demanda pode ser desempenhado através de aumento de produção ou administração das demandas dos sistemas.

Em se tratando da administração das demandas dos sistemas, fala-se especificamente das perdas de água tratada nos sistemas de abastecimento público. A Secretaria Nacional de Saneamento Ambiental (2008) aponta que as perdas são um dos principais indicadores de desempenho de uma prestadora de serviços de saneamento em todo o mundo, e que esta componente pode ser encontrada desde a captação até a distribuição, em diferentes dimensões em cada sistema.

Dando suporte a esta ideia, o conceito de perdas é citado pela ABNT na norma ISO 24510 (2012) em dois pontos importantes da avaliação da eficiência das companhias de abastecimento público: os serviços prestados aos usuários, no que tange a quantidade de água

fornecida aos usuários, e da sua postura de proteção ao meio ambiente, referindo-se ao uso sustentável dos recursos naturais.

Em 2016, através da Lei 13.329/2016, foi acrescentado no texto da Lei 11.445/2007, que estabelece as diretrizes nacionais para o saneamento básico, a instituição de regime especial de incentivos para investimentos que visem à redução de perdas de água. (BRASIL, 2016)

A ABES – Associação Brasileira de Engenharia Sanitária, alerta para a necessidade de haver uma cultura de combate a perdas nas companhias de saneamento, sob pena de que se nada for feito, os volumes perdidos crescem naturalmente seja por perdas aparentes (fraudes, desgaste nos hidrômetros) ou perdas reais (deterioração das tubulações e vazamentos nas redes e ramais) (2015).

Mutikanga, Sharma e Vairavamoorthy (2009) são incisivos em delimitar o papel fundamental das concessionárias quanto a necessidade de administrar de forma responsável o recurso, e em afirmar que em casos de cidades onde há escassez hídrica, com racionamento, a ocorrência contínua de vazamentos é praticamente um comportamento criminoso.

Do ponto de vista energético, Colombo e Karney (2002) afirmam que as perdas demandam mais energia para que seja possível manter níveis de serviços aceitáveis. Os autores Souza, Covas e Soares (2010) vão ainda mais longe, apontando que a melhoria da eficiência energética dos sistemas de abastecimento envolve a redução do consumo de energia, e que, por esta razão, o desenvolvimento e implantação de estratégias efetivas de redução de perdas é tão importante para as companhias de saneamento, “dada a sua responsabilidade por atingir os anseios dos consumidores, garantindo sustentabilidade técnica e econômica das infraestruturas, protegendo o meio ambiente e promovendo o desenvolvimento sustentável da comunidade”.

1.1.3 Problema central da pesquisa

O problema identificado é que ligações de água fraudadas, ou seja, que de alguma forma consomem água de maneira não autorizada, seja através de intervenções sobre os equipamentos de micromedição (hidrômetros), ou através de by-pass, geram custos financeiros e ambientais para a sociedade e companhias de saneamento.

Estes custos são incrementados a medida em que, quando não se paga pelo consumo os consumidores tendem a utilizar a água de forma irracional, sem se preocupar com desperdícios.

1.1.4 Hipóteses da pesquisa

A hipótese sobre a qual se apoia a pesquisa delimitada por este estudo é de que é possível traçar, através de séries temporais, ou seja, dados históricos de consumo das ligações de uma determinada localidade, um perfil que seja comum a parcela de consumidores que praticam o consumo não autorizado. Através da análise destes dados históricos, podem ser criados mecanismos automatizados de identificação de ligações fraudadas, e assim combater esta prática.

1.1.5 Justificativas da pesquisa

O trabalho desenvolvido contribuirá para um uso mais eficiente da água, a partir do momento em que um modelo de identificação mais célere de ligações de água que tenham sido fraudadas permite às companhias tomar medidas de forma imediata, reduzindo o uso ineficiente do recurso natural e ao mesmo tempo o consumo de energia necessário para produzir e transportar a água até a ligação em questão.

Dessa forma, conhecer essas variáveis e analisa-las se justifica na necessidade de conhecer os mecanismos que oneram os custos de energia e o seu consumo por parte das companhias de Saneamento. Estabelecer tais conexões faz com que seja possível adotar medidas para combater tais usos ineficientes de água e, conseqüentemente, energia, reduzindo o uso dos recursos naturais utilizados para sua produção e conseqüentes impactos sobre o meio ambiente.

1.2 OBJETIVOS

1.2.1 Objetivo Geral

O objetivo das ações propostas é criar de um modelo de trabalho que incorpore a inteligência artificial dos algoritmos de máquinas de vetor de suporte (SVM) e que se transforme em um produto a ser utilizado pelas companhias de saneamento responsáveis por serviços de abastecimento de água tratada, para que possam tornar mais eficiente a identificação de perdas aparentes em ligações dos sistemas que operam, reduzindo o mau uso dos recursos

necessários para a produção de água tratada, dentre eles a energia, tornando-as mais eficientes do ponto de vista econômico, social e ambiental.

1.2.2 Objetivos Específicos

Para atingir o objetivo proposto, este trabalho delimitou os seguintes objetivos específicos:

- Aplicar os modelos estudados na determinação de uma assinatura padrão para identificação de ligações com ocorrência de consumo não autorizado de água tratada (fraudes);
- Desenvolver um sistema capaz de apontar ligações de um sistema de abastecimento de água que apresentem fraudes;
- Avaliar a eficiência do modelo proposto junto a um sistema de abastecimento público de água.

2 REVISÃO BIBLIOGRÁFICA

2.1 Eficiência Energética em Sistemas de Saneamento

O setor de saneamento, segundo Gomes e De Carvalho (2009), é o mais estratégico no que diz respeito ao uso conjunto de água e energia em tempos que o uso racional destes elementos é indispensável para o desenvolvimento econômico e social, pela necessidade de preservação dos recursos naturais. O autor afirma também que o valor destes insumos é cada vez maior, e que sua racionalização evita a oneração dos custos de produção do setor.

Apesar do enorme potencial estratégico, infelizmente o setor de saneamento, de acordo com Martins (1999) por sua estruturação regionalizada, com base em companhias estaduais e serviços municipais, não possui suporte em nível nacional que permita a existência de modelagem técnico-operacional capaz de induzir a disseminação das tecnologias aplicáveis ao segmento. Para o autor, dessa forma, a eficiência energética no setor de saneamento precisa passar por mudanças estruturais, considerando-se o atraso tecnológico decorrente da falta de investimentos.

Em 1985, o governo Federal, através do Ministério de Minas e Energia e Eletrobrás, criou o Programa Nacional de Conservação de Energia Elétrica (PROCEL). O programa foi criado com o objetivo de promover o uso eficiente da energia elétrica e combater o seu desperdício, com o desenvolvimento de conhecimento sobre o consumo eficiente da energia e fomentando as práticas com esta finalidade. Além disso, postergar os investimentos no setor elétrico como aumento de demanda, infraestrutura de carga, mitigando assim os impactos ambientais e colaborando para um Brasil mais sustentável.

Com o mesmo intuito, em 2003, a ELETROBRAS/PROCEL instituiu o PROCEL SANEAR – Programa de Eficiência Energética em Saneamento Ambiental, buscando o uso eficiente de energia elétrica nos conjuntos moto-bombas dos sistemas de saneamento e também ações quanto à conservação da água, visando integrar os dois temas e o desenvolvimento e a operacionalização de projetos e políticas governamentais articuladas (PROCEL, 2015).

A meta do programa PROCEL SANEAR é reduzir em 15% o desperdício de energia elétrica no setor, com a adoção de ações que visem a modulação de carga em relação à ponta dos sistemas elétricos, o controle das vazões de recalque em relação à demanda da rede de distribuição, o dimensionamento adequado dos equipamentos eletromecânicos e a automação operacional de sistemas com gerenciamento e supervisão em tempo real (Martins, 1999).

2.2 Redução de Custos com Energia em Saneamento

Aubuchom e Roberson (2014), discorrem sobre o termo “energia incorporada”, que de acordo com os autores, trata da quantidade total de energia usada para produzir e fornecer uma determinada quantidade de água. Essa energia varia de concessionária para concessionária, em relação ao tipo de fonte de geração de energia, os processos de tratamento adotados e outros fatores relacionados a distribuição e uso (JOHNSTON e KARANFIL, 2013, apud AUBUCHOM e ROBERSON, 2014).

Young (2013) também discorre sobre o tema, afirmando que água e energia estão ligadas de forma inerente, e que esta forma de interseção é conhecida como “nexo água e energia”. A partir do estabelecimento desta relação, a autora afirma que programas de eficiência que buscam reduzir desperdício de água têm ação direta sobre a redução do uso de energia (e vice e versa). Essa relação fica evidente quando se observa a necessidade de energia para transportar, tratar, aquecer, resfriar, e reciclar a água nas companhias (YOUNG e MACKRES, 2013).

Da mesma forma no Brasil, a disponibilidade de água e a geração de energia estão intimamente ligadas. Em tempos de crise hídrica, como ocorrida recentemente em 2015, a energia consumida precisa ser suprida por usinas termoeletricas, que utilizam combustíveis fósseis (gás natural, carvão mineral, óleo diesel, óleo combustível) para operação, o que não apenas aumenta o custo da geração como também implica em mais emissões de carbono que afetam o clima do planeta.

De acordo com Tsutiya (1997), o fim gradativo do subsídio que era concedido pelo Governo nas tarifas de energia elétrica para os serviços públicos de abastecimento de água e disposição de esgotos, o aumento das tarifas dessa energia acima dos índices inflacionários e a cobrança do Imposto sobre Operações relativas à Circulação de Mercadorias e sobre Prestação de Serviços de Transporte Interestadual e Intermunicipal e de Comunicações (ICMS), causaram o aumento de despesas de energia elétrica. Nota-se que além dos custos financeiros do consumo de energia, deve-se considerar também os custos ambientais.

O Green Peace (2010) cita o Painel Intergovernamental de Mudanças Climáticas (IPCC), órgão das Nações Unidas, que afirma que a temperatura mundial pode aumentar – em um cenário extremo – em até 6,4°C até o final deste século, causando a maior alteração climática já vivida pelo ser humano. Para evitar que isso ocorra, uma política global deve ter por objetivo manter esse aumento em menos de 2°C, ou mesmo 1,5°C, em relação à temperatura registrada antes da revolução Industrial, no século 18. Uma das medidas apontadas pelo Green Peace, é

a necessidade de impor padrões de eficiência energética para equipamentos elétricos. Como resultado dessas ações haveria redução de demanda de energia e muitas usinas termoeletricas poderiam ser desativadas e possibilitaria a redução de emissões de gases do efeito estufa na atmosfera.

Tsutiya (1997) descreve as diversas formas de se reduzir os custos com energia em sistemas de saneamento, separando-as em três grupos.

a) Redução de custo sem diminuição do consumo de energia elétrica:

O autor divide o primeiro em Redução de Custo sem investimento, quando existem situações que independem de investimento em novas instalações ou equipamentos e que são solucionados através de correspondência encaminhada às concessionárias; e Redução do Custo com Investimento, que depende de recursos investidos em equipamentos e instalações, após a realização de pequenos estudos. Como se trata de investimentos não muito elevados, são bastante utilizados. Entretanto, as medidas que não demandam investimentos, e tratam de gestão, são as mais utilizadas para a redução de custos.

b) Redução do custo pela diminuição do consumo de energia elétrica:

Neste caso o autor afirma que, tendo em vista que o custo do consumo de energia elétrica é o produto da energia consumida em kWh pela tarifa do kWh, é necessário reduzir a altura manométrica ou o volume da água bombeada, ou aumentar a eficiência dos conjuntos motor-bomba para diminuir os custos de energia.

c) Redução de custo pela alteração do sistema operacional das estações elevatórias:

Tratam-se de adequação dos conceitos adotados durante a concepção de um determinado sistema de elevatórias, uma reengenharia que possibilita a redução dos custos com energia e, em grande parte dos casos, a adequação de pressões na rede a valores mais constantes.

O projeto aqui proposto visa discorrer sobre ações que as companhias podem tomar sobre o volume disponibilizado e através delas reduzir a quantidade de energia demandada pelos sistemas produtores. É importante frisar que o uso mais racional da água também otimiza o uso das infraestruturas existentes, adiando ou evitando a necessidade de investimentos com a expansão dos sistemas (LOUREIRO, 2010).

2.3 Perdas de Água em Sistemas de Abastecimento Público

De acordo com Jansen e Schulz (2006), a água vem se tornando mais escassa de forma crescente, em decorrência do crescimento populacional e econômico e, conseqüentemente, a urbanização, além de mudanças climáticas. Os autores citados afirmam que este problema demanda prioritariamente administrar este recurso com eficiência, especialmente em países em desenvolvimento.

Moreira (2011) aponta que é papel das empresas de saneamento garantir a eficiência na distribuição e na qualidade da água oferecida. O autor afirma que para esta finalidade, os administradores dos serviços criaram sistemas de distribuição muito complexos, entretanto resta muito a se fazer em relação a gestão da água. Nesse aspecto, a prioridade é a gestão de perdas de água nas redes de distribuição.

Garantir um sistema de abastecimento de água adequado está se tornando um desafio crescente em muitos países no mundo todo (MUTIKANGA, SHARMA E VAIRAVAMOORTHY, 2009).

A Secretaria Nacional de Saneamento Ambiental (2008) aponta que as perdas são um dos principais indicadores de desempenho de uma prestadora de serviços de saneamento em todo o mundo, e que esta componente pode ser encontrada desde a captação até a distribuição, em diferentes dimensões em cada sistema.

As perdas, além do ponto de vista da eficiência do processo, ambiental e social, devido ao uso não racional da água, também se trata de um problema econômico, “pois têm um forte peso nas despesas de uma empresa gestora de água” (MOREIRA, 2011).

De acordo com a Alliance to Save Energy (2012), uma entidade composta por diversos envolvidos da indústria americana que promovem a eficiência energética, para tornar mais eficiente o uso da água as autoridades envolvidas devem entender a relação entre água e energia, que estão diretamente relacionadas e unidas de forma que “Cada litro de água que se move pelo sistema representa um significativo custo de energia.”. A entidade enfatiza que vazamentos, furtos, desperdício e ineficiência na distribuição impactam a demanda de energia do setor diretamente.

Sob um olhar mais atento, é possível entender mais as perdas de água tratada considerando o tema através do critério utilizado pela IWA – International Water Association, que as divide em perdas reais e perdas aparentes (Lambert, 2003 apud XIN et al., 2014).

Ainda de acordo com a ABES (2015), essa caracterização das perdas reais e aparentes permite uma diferente abordagem econômica para cada tipo de perda. Sobre as “perdas reais” recaem os custos de produção e distribuição da água, e sobre as “perdas aparentes”, os custos de venda da água no varejo, acrescidos dos eventuais custos da coleta de esgotos.

2.3.1 Perdas reais de água

2.3.1.1 Conceito

Oliveira (2013) inicia a descrição do tema através de referência a IWA (International Water Association). Perdas reais é a nomenclatura definida pela IWA para as perdas físicas, ou seja, a água produzida e transportada “que não chega a ser consumida pelo consumidor final”.

A ABES, em seu guia de contribuições técnicas sobre o controle e redução de perdas nos sistemas públicos de abastecimento de água, trata as perdas físicas como as ocorrências em que o volume de água disponibilizado no sistema de distribuição pelas operadoras de água não é utilizado pelos clientes, sendo desperdiçado antes de chegar às unidades de consumo.

As perdas reais tratam-se de vazamentos (visíveis ou invisíveis), seja na rede de distribuição ou nos ramais domiciliares (a montante do hidrômetro), ou extravasamentos de reservatórios (ABES, 2015).

De acordo com Gonçalves (2013), dentre os casos citados acima, as perdas reais estão associadas principalmente em função de vazamentos, com maior concentração na rede de distribuição, em particular com grande recorrência nos ramais prediais.

Cardoso (2013) reforça o conceito através da citação de que perdas reais resultam de fissuras, rompimentos ou extravasamentos, desde a captação até os “pontos de consumo” decorrentes do fato do sistema de abastecimento não ser totalmente estanque.

Tavares et al. (2015) detalham os tipos de vazamento em quatro tipos: Vazamentos visíveis, Vazamentos invisíveis, Vazamentos inerentes e Extravasamento de reservatórios:

Vazamentos Visíveis

Vazamentos visíveis são facilmente detectáveis, uma vez que afloram à superfície. As demandas são trazidas pela população ou pela equipe de manutenção. Para todos os vazamentos são abertas ordens de serviços, cuja

execução é acompanhada por equipe técnica. Para diminuir a incidência desse tipo de vazamento recomenda-se a redução da pressão nas redes de água. A agilidade na detecção e reparo desse tipo de vazamento é primordial para um controle de perdas eficiente.

Vazamentos Não Visíveis

Os vazamentos não visíveis são aqueles que não afloram à sua superfície, o que torna difícil a detecção. Para tanto, são realizadas pesquisas com a utilização de equipamentos amplificadores de som e hastes de escutas. As pesquisas deste tipo de vazamento são executadas através da análise minuciosa da rede de distribuição. Em alguns casos são mais eficientes no período noturno, uma vez que a interferência sonora é menor. Vale ressaltar que a redução de pressão também tem efeito preventivo nesses casos.

Vazamentos Inerentes

Vazamentos inerentes são aqueles que aparecem nas juntas e conexões utilizadas na infraestrutura de distribuição. Esses, na maioria das vezes, não são detectáveis de forma visível ou através equipamentos de escuta. Para combater essas perdas é necessária, além da redução de pressão, a utilização de materiais de qualidade e adequados às condições do projeto.

Extravasamento de Reservatórios

Diminuir a ocorrência de extravasamento de água em reservatórios é de extrema importância para a redução de perdas, uma vez que o volume desperdiçado pode exceder o economizado com a retirada de vazamentos. Para controlar este tipo de perda, é importante instalar sistemas de controle de nível automatizados eficientes, ou sistema de monitoramento remoto que indica a quantidade de água existente em cada equipamento, além de controlar o acionamento das bombas do sistema de distribuição.

(TAVARES et al., 2015).

Cardoso (2013) reafirma esta subdivisão, citando Farley et al. (2008), em consonância com Oliveira (2013), quando todos tratam as perdas físicas/reais em um nível inicial em vazamentos em redes de distribuição e ramais domiciliares e também extravasamentos de reservatórios. Em seguida, os autores detalham os vazamentos de rede e ramal em três tipos:

facilmente detectáveis (visíveis), dificilmente detectáveis (invisíveis, necessitam de equipamentos de detecção acústica) e as perdas de base (inerentes).

É importante destacar que, nos últimos anos, muito trabalho foi dedicado com foco na componente das perdas reais do balanço hídrico (Lambert et al., 1999; Fanner et al., 2007 apud XIN et al., 2014). Isso se deve pelo fato das perdas reais serem mais fáceis de investigar quando comparadas com as perdas aparentes. Assim, pesquisadores deram mais atenção as perdas físicas e controle de vazamentos.

2.3.1.1. Efeitos das perdas reais

Oliveira (2013) ressalta os inúmeros efeitos causados nos sistemas de abastecimento, pelos vazamentos, tanto para as companhias quanto para os clientes e o meio ambiente. O aspecto mais relevante, de acordo com Farley et al. (2008, apud CARDOSO, 2013) é financeiro. Tal afirmação se apoia sobre questões financeiras relativas à reparação das tubulações, encargos diretos relativos ao volume de água perdido, custos de produção excessivos (energia, insumos de tratamento, mão de obra), erros de medição em decorrência da variação de pressão pela entrada de ar na rede (e consequente falha no faturamento), pagamento de compensações aos consumidores, indústrias e comércio, custos para lidar com reclamações, além de danos à imagem da empresa junto a população.

Com a redução das perdas físicas, a empresa pode produzir uma quantidade menor de água para abastecer a mesma quantidade de pessoas, o que tem como consequência a redução dos custos com diversos itens, tais como (ABES, 2015):

- Produtos químicos;
- Energia elétrica;
- Compra de água bruta (nos casos em que há cobrança pelo uso da água);
- Mão de obra.

2.3.2 Perdas de água aparentes

Dentro do tema das perdas de água e consequente demanda de energia de forma desnecessária, aborda-se o problema das perdas aparentes.

Ao contrário das perdas reais, que tratam de volume que não é consumido, as perdas aparentes se tratam de volume que é consumido, porém não faturado (XIN et al., 2014). Este termo se refere à porção da água disponibilizada que não é faturada pelas companhias, em

decorrência de erros de medição (macro e micro), fraudes por partes dos consumidores, consumo clandestino e até falhas no cadastro comercial (VICENTINI, 2012). Lambert et al. (1999, apud XIN et al., 2014) também definem este tipo de fraudes como uso não autorizado, decorrente de furto de água e conexões ilegais nas redes de distribuição de água.

Loureiro (2010) descreve o problema das perdas aparentes como a inclusão de “todas as imprecisões associadas às medições da água produzida e da água consumida, assim como o consumo não-autorizado (por furto ou uso ilícito)”. Entretanto, a autora vai além das demais referências citadas, colocando ao lado destes dois aspectos também os erros de transferência de dados entre o contador e o sistema de leitura e faturamento, bem como os erros decorrentes da análise de dados armazenados nos sistemas comerciais.

Ainda assim, Rizzo e Cilia (2005, apud XIN et al. 2014) demonstraram que a imprecisão de medição dos hidrômetros das ligações, causadas por submedição em baixas vazões ou adoção de medidores mal dimensionados para as classes de vazão em que trabalham, são a principal razão das perdas de água nos sistemas de abastecimento, e por isso, melhorar a precisão dos medidores é a estratégia chave para controlar as perdas aparentes.

Vicentini (2012) salienta que a redução deste tipo de perda tem impacto direto sobre a redução de desperdício por parte dos consumidores, ou seja, é capaz de reduzir o uso de energia necessária para bombear esta água até o ponto de consumo, afinal, além das imprecisões nas medições de água produzida e consumida, grande parte deste volume é desperdiçado por usuários que praticam consumos não autorizados por furto ou uso ilícito.

Diante destas afirmações, é possível compreender a necessidade de se investir no combate as perdas aparentes. Mutikanga, Sharma e Vairavamoorthy (2009), em seu trabalho sobre perdas de água em países em desenvolvimento, observa que são necessárias mais pesquisas sobre as perdas aparentes, especialmente nestes países onde as mesmas são mais proeminentes, de forma que possam chegar ao mesmo patamar das intervenções sobre as perdas reais.

No caso estudado, através do balanço hídrico, é possível afirmar que as perdas aparentes representam em volume uma proporção de 22% do volume total das perdas. Entretanto, do ponto de vista financeiro, as perdas aparentes praticamente se igualam ao valor das perdas reais, atingindo um montante estimado em R\$ 11,5 milhões aproximadamente, ou seja, 46,2% do valor total das perdas na localidade no ano de 2016. Neste cenário, as perdas causadas por

consumo não autorizado neste sistema no mesmo ano foram estimadas em mais de R\$ 2,8 milhões.

No caso das perdas aparentes, a estratégia assenta essencialmente na análise do balanço custo-benefício entre o investimento necessário para as reduzir e os benefícios financeiros daí decorrentes (LOUREIRO, 2010).

De maneira geral, os programas de redução de perda de água comercial são financeiramente atraentes, pois geram um retorno financeiro rápido. Com a redução das perdas aparentes, decorrentes de fraudes nas ligações, consumo não faturado, falta de hidrômetros, problemas de medição, dentre outros, a principal consequência é o aumento do volume faturado e, conseqüentemente, da receita (ABES, 2015).

2.3.2.1 Consumo não autorizado de água tratada

Prestadores de serviço em todo o mundo passam por problemas com a desonestidade de seus clientes ou consumidores, sejam operadoras de cartões de crédito, sistemas de distribuição de energia elétrica, distribuição de água ou outros.

É possível encontrar vasta literatura técnica sobre o tema principalmente em relação as concessionárias de energia, referente ao que chamam de Perdas Não Técnicas (*NTL – Non Technical Losses*). Os técnicos que estudam o assunto apontam que a desonestidade dos consumidores de eletricidade é um problema comum enfrentado pelas companhias, e que encontrar medidas eficientes para detectar o consumo não autorizado tem sido uma área de pesquisa ativa nos últimos anos (NAGI et al., 2010).

Spirića, Dočićb e Stankovićb (2015) afirmam que esse tipo de comportamento pode ser realizado por clientes ilegais, não registrados, que dessa forma não passam de consumidores, não podendo ser chamados de clientes. Todavia, também podemos observar tal fato em pontos de medição legalizados, em clientes cadastrados junto as concessionárias. De acordo com os autores, este problema é mais comum em países em desenvolvimento ou em países com uma renda média per capita baixa.

Nagi et al. relatam que, devido aos problemas de fraude, as concessionárias perdem grandes quantidades de dinheiro. Definem ainda a fraude como “uso ilegal ou desonesto de energia elétrica com a intenção de evitar cobrança”. Por fim apontam a dificuldade para detectar o comportamento negativo entre os consumidores honestos, e que as concessionárias nunca

serão capazes de eliminar totalmente essas ações, entretanto podem encontrar medidas para detectar, prevenir e reduzi-las.

O Brasil se enquadra na descrição acima, e as companhias de energia também possuem grande dificuldade em identificar e coibir as fraudes. Henriques et al. (2014) afirma que no país são comuns as “*conexões ilegais*” e que a energia realmente consumida por elas é maior do que a registrada nos medidores. O autor também relata que as perdas representam prejuízo financeiro para as concessionárias e que os consumidores honestos acabam sendo penalizados, considerando que recentemente as perdas por fraude foram inseridas nos preços da eletricidade. Por fim, relata que a busca por medidas para mitigar as fraudes é comum em setores diferentes como a telecomunicação ou sistemas de distribuição (água, gás, coleta de esgoto, etc.).

Aqui é possível traçarmos um paralelo com os sistemas de distribuição de água, onde os problemas enfrentados são os mesmos. Os consumidores que atuam dessa forma representam perdas de faturamento para as companhias de saneamento, um valor extremamente difícil de ser mensurado, considerando que o consumo de uma ligação fraudada ou clandestina não obedece aos padrões de um cliente que está sendo cobrado.

Fetterman et al. (2015) fundamentam tal afirmação quando descrevem que tal dificuldade “*resulta em prejuízos financeiros para a concessionária em razão do volume de água não faturado*”. Tal dificuldade surge em razão das diversas fontes de variabilidade existente no padrão de consumo de uma unidade, tais como alteração na quantidade de moradores, viagens, clima, vazamentos entre outros. Por isso é possível afirmar que a detecção das fraudes por parte das companhias de distribuição de água tratada se constitui em importante fator de economia.

A ABES (2015) aponta ainda que, ao fraudar uma ligação, e perceber que isso não suscitou nenhuma reação da operadora de saneamento, o cliente se sente incentivado a sugerir a outros clientes um procedimento semelhante, aumentando as perdas por fraudes. Gonçalves (2009) afirma que estas ocorrências podem ser afetadas por fatores “*sociais e culturais, influências política, financeiras, institucionais e organizacionais*”.

Em seu estudo sobre as perdas não técnicas de energia, Nagi et al. (2010) afirmam que as técnicas atuais de identificação, ou seja, inspeções nas instalações dos consumidores, são conduzidas sem nenhum foco específico, em decorrência da inexistência de um sistema que possa reduzir as ligações suspeitas de fraude de forma eficiente. Os autores citam a companhia

de energia da Malásia que tem taxas de detecção de fraudes na ordem de 3% sobre as conexões pesquisadas.

Spirića, Dočićb e Stankovićb (2015) também vão ao encontro das informações obtidas no estudo da Malásia, afirmando que as inspeções nos pontos de medição dos consumidores, geralmente, são despreparadas e sem sofisticação, com baixos índices de descoberta de fraudes.

Nas concessionárias de distribuição de água observa-se os mesmos fatos. Gumier e Luvizotto Junior (2007, APUD FETTERMAN ET AL., 2015) afirmam que “o *principal método de detecção utilizado é a inspeção com a utilização de equipamentos acústicos, um método vagaroso e dispendioso*”. O artigo faz ainda uma crítica sobre a determinação dos pontos pesquisados, que na maioria dos casos, é realizada ao acaso.

2.4 Mineração de Dados e a Identificação de Perdas Aparentes

Diante o exposto nos itens anteriores, sugere-se aprofundar-se no assunto de detecção de fraudes em ligações nos sistemas de distribuição de água através do uso de Mineração de Dados, com base em dados históricos de consumo dos clientes, visando aumentar a eficiência na busca pelas companhias de saneamento.

A redução do desperdício de água tratada, com objetivo de atender metas de sustentabilidade é cobrada diretamente das concessionárias responsáveis pelo seu fornecimento. As formas de atingir tais objetivos incluem o desenvolvimento de fontes alternativas de água, reúso de água e redução de consumo (CARDELL-OLIVER, 2013).

O quesito redução de consumo engloba muito bem a questão das perdas aparentes por fraudes ou consumos não autorizados. Para encontrar respostas acerca de indícios que apontem para características comuns a grupos de ligações em que existem perdas aparentes, é necessário um método de avaliação e correlação. A proposta é a utilização da metodologia conhecida como Mineração de Dados, ou mineração de dados. Isso se faz necessário pois ferramentas eficientes de suporte a tomada de decisão e abordagens para análises de sistemas são necessárias para aplicar esforços sobre o tema da melhor forma possível (HIGHTOWER, 2011).

Em seu estudo sobre mineração de dados em bancos de dados, Silva e Antunes (2015) apontam como o modelo de armazenamento de informação mudou nos últimos anos, tornando mais simples a análise das mesmas, apesar de dois grandes desafios primários: a grande quantidade de dados e as diversas dimensões de análise simultâneas. As técnicas de Mineração

de Dados têm a função de aproveitar essa grande massa de informações e superar os desafios citados.

Shmueli et al. (2016) se refere ao termo Mineração de Dados como métodos de análise de negócios que vão além de cálculos, técnicas descritivas, relatórios e outros métodos baseados em regras de negócio. Resume seu trabalho sobre o assunto em métodos estatísticos e aprendizado por computadores que dão suporte a tomada de decisão. Por fim, uma variedade de técnicas para explorar dados e construir modelos de forma mais fácil que na estatística clássica. O autor cita a descrição de Dary Pregibon's do tema como "Estatística em escala e velocidade" (PREGIBON, 1999, apud SHMUELI ET AL., 2016).

Rathod e Garg (2016) apresentam o termo "Mineração de Dados" como a técnica utilizada para extrair "informações preditivas ocultas" em grandes bancos de dados encontrados em qualquer aplicação da área científica ou de engenharia. Em seu trabalho, apontam como a técnica tem sido utilizada na área de distribuição de energia para previsão de demanda, determinação de planos de tarifas e, que tem como objeto deste trabalho podendo ser utilizada de maneira análoga na área de abastecimento de água, a classificação de tipos de consumidores.

Assim como apresentado na área da energia, Bakirli et al. (2012) propõem um estudo que aponta soluções em Mineração de Dados para concessionárias locais de serviços de forma a tornar os seus mecanismos de tomada de decisão mais simples. Para atingir o objetivo traçado, os autores desenvolvem uma forma de estimarem atividades futuras através de dados passados.

Para os autores da Turquia, Mineração de Dados extrai padrões escondidos em grandes conjuntos de dados para uma forma compreensível ao ponto de vista humano. Dessa forma é possível apontar fatos importantes que eram desconhecidos e saber o que vem a frente.

Da mesma forma, Wang, Cardell-Oliver e Liu (2016) afirmam que dados temporais obtidos através de hidrômetros inteligentes podem auxiliar nas estratégias de definição de tarifa, intervenções na conservação de recursos e planejamento de infraestrutura. Uma grande quantidade de dados de consumo também depende de uma análise eficiente para caracterização dos comportamentos dos usuários (padrões). O uso de Mineração de Dados tem sido utilizado para esta análise de forma automática, evitando análises manuais que consomem tempo em excesso onerando a eficiência do processo de investigação.

Em estudo sobre o consumo de água encanada em sistemas de abastecimento rural em Gana, Kulinkina et al. (2016) afirmam a importância da demanda contínua e a geração de receitas para garantir a manutenção dos serviços de abastecimento através do tempo, e relatam

as suas descobertas sobre a análise dos dados de consumo nos hidrômetros existentes: métodos qualitativos como forma de permitir um planejamento mais eficiente e desenvolvimento sustentável para os administradores destes sistemas.

Fetterman et al. (2015) realizaram estudo em que, a partir da dificuldade em determinar os pontos de pesquisa sobre a perdas aparentes (fraudes), buscam apontar de forma mais assertiva eventuais ligações que possam estar fraudadas, ações que o autor afirma que já são existentes junto a operadoras de cartão de crédito, por exemplo, e em concessionárias de serviços públicos de forma mais adiantada junto ao setor de energia.

Shmueli et al. (2016) tratam o que chamam de analítica preditiva também como um método amplamente utilizado, citando o sistema de serviço interno de arrecadação americano que descobriu como aumentar em 25 vezes a assertividade na busca por evasões de impostos a partir de ações baseadas em tais modelos preditivos gerados a partir de análise de dados.

Para Rathod e Garg (2016), de acordo com o tipo de dados e o objetivo desejado, o método de mineração de dados se aproveita de diferentes abordagens e modelos. De forma geral, classificam os métodos em preditivo e descritivo, que são técnicas de associação ou regras de associação e clustering, respectivamente. Esses métodos podem explorar dados baseados em mineração em texto ou na web, áudio e vídeo (multimídia mining), sensoriamento remoto e aplicações SIG (mineração de dados espacial) e séries de dados temporais.

Bakirli et al. (2012) afirmam que os principais métodos de Mineração de Dados são Classificação, Clusterização, aprendizado de regra de associação ou valores atípicos (outliers).

Em seu estudo, Fettermann et al. (2015) apresentam três formas de análise de dados disponíveis em sistemas de água para identificar eventuais problemas relacionados a perdas de água (Vazamentos ou fraudes). Dentre eles, a previsão de demanda, com base na qual é possível identificar uma possível fraude através de um desvio acima do aceitável em um valor de consumo esperado, e a identificação de dados atípicos (outliers), ou seja, observações que se desviam demasiado das demais pertencentes ao grupo.

Através da utilização dos conceitos acima de extração de informações em bancos de dados, Mineração de Dados, é possível agrupar indivíduos a partir de características comuns que os mesmos tenham dentro de um universo de comportamentos observados. Estes agrupamentos, podem ser feitos através da identificação de comportamentos específicos de uso da água, ou padrões de consumo.

Um padrão é uma estrutura dentro de um banco de informações, que pode ser definida de diversas formas. Entretanto, o conceito utilizado normalmente é o de padrões frequentes, ou seja, padrões que ocorrem repetidas vezes entre os dados, e de acordo com Mehta e Dang (2011), grande parte da literatura sobre o tema se propõe em formular padrões úteis e desenvolver algoritmos eficientes para descobrir tais padrões.

Silva e Antunes (2015) relatam que existem muitos algoritmos criados para encontrar padrões frequentes em grandes bases de dados.

Cardell-Oliver (2013) define Padrões de consumo como *blocos (subconjuntos) de leituras de hidrômetros especificadas por padrões de volumes e datas no calendário*. Os autores são bastante seguros em afirmar que estas “assinaturas”, a forma como se referem aos padrões de consumo, permitem identificar comportamentos que são alvos promissores para a conservação de água. Além disso, a descoberta destes padrões pode ser automatizada com uso de programas eficientes de computador.

Para Cardell-Oliver, Wang e Gigney (2016), “*conceber estratégias de negócios para gerenciar consumo humano de recursos, como a água ou a eletricidade exige conhecimento de quem, quando e como esse recurso está sendo utilizado*”.

Um exemplo interessante é o do estudo realizado por Kulinkina et al. (2016), que a partir de dados de hidrômetros instalados em sistema rural de abastecimento na região leste de Gana, identifica a relação de padrões de consumo a partir de dados de consumo registrados ao longo do tempo, relacionados com dados geoespaciais e informações de chuva na região.

Em outro caso, uma classe de clientes com series anômalas de consumo é formada baseada em resultados de fraude anteriores. Em seguida busca-se na base de dados clientes que se enquadrem nessa classe e que podem pertencer a um grupo de ligações fraudadas, ou seja, considerados suspeitos. Essa relação de ligações gera uma lista que irá orientar a busca por fraudes de forma mais assertiva. O resultado obtido no estudo indica que em 93,4% das séries de dados estudadas existe um fator não aleatório, ou seja, regras específicas para grupos de consumidores (SPIRÍČA, DOČÍČB, STANKOVIČB, 2015).

De forma semelhante, na Malásia, Nagi et al. (2010) encontraram uma lista de 102 suspeitos de fraude em um grupo de 10.000 consumidores.

Já no trabalho realizado por Henriques et al. (2014), apenas 188 dos 10 279 clientes (menos de 2%) dos registros selecionados para mineração foram apontados para inspeção. Entretanto, a taxa de identificação correta de fraude (entre 7% e 20%) melhorou

significativamente as campanhas anteriores de detecção da companhia, com especial presença de fraudes observadas em clientes de consumo médio-alto (HENRIQUES et al., 2014).

2.4.1 Algoritmos de Classificação

Um classificador é uma função que atribui classes ou rótulos a objetos de um grupo usando estruturas de dados internas para compará-los com demais componentes ou comportamentos previamente identificados. Essa identificação prévia é realizada a partir de um indutor, um algoritmo de análise que constrói o classificador a partir de um dado conjunto de dados e gera o algoritmo classificador.

Dessa forma, a precisão de um classificador induzido por algoritmos de aprendizagem supervisionados é importante não só para prever sua precisão de previsão futura, como também para escolher o classificador adequado (KOHAVI, 1995).

A necessidade de determinar a precisão destes algoritmos, e consequentemente o seu sucesso não envolve apenas a definição de técnicas de classificação adequadas de atributos de qualidade, mas também o estabelecimento de métodos para a avaliação dos mesmos (GOULAO e BRITO E ABREU, 2005).

Para tal avaliação, a literatura analisada aponta os métodos de validação cruzada como forma amplamente empregada. Através desta metodologia é possível avaliar a precisão para um determinado modelo na prática, ou seja, o seu desempenho para um novo conjunto de dados.

A análise desempenhada pode ser aplicada em dois momentos diferentes do processo de avaliação de precisão:

- consistência interna – Avalia a capacidade do classificador de produzir saídas corretas para entradas pertencentes à amostra usada na construção do modelo.
- consistência externa – Avalia a capacidade do classificador de produzir saídas corretas para entradas da mesma população, mas não pertencentes à amostra usada na sua criação.

2.4.1.1 Treinamento e validação

Em resumo, validação cruzada trata da divisão de um conjunto de dados de estudo em subconjuntos distintos, antes de apresentá-los ao classificador. Os dados são separados em dois conjuntos principais: conjunto de treino (*training set*) e conjunto de teste (*test set*). No caso deste projeto, o *training set* é o conjunto que é apresentado para o classificador na etapa de treinamento, onde o algoritmo vai aprender os padrões importantes para discriminar classes

positivas (fraudes) de negativas (não-fraudes). Nessa etapa os rótulos de cada exemplo são conhecidos.

Após treinar o classificador no *training set*, aplica-se o conhecimento aprendido em uma nova base nunca vista antes: *test set*. O classificador efetua as predições para o *test set* e avalia sua performance comparando o que foi predito e os rótulos reais de teste.

De acordo com Reitermanová (2010), muitas vezes, um algoritmo classificador pode apresentar resultados de generalização fraca, o que é decorrente de excesso de treinamento. O autor propõe como um método comum para evitar o excesso de treinamento a validação cruzada de retenção, *Hold-out*.

A validação cruzada hold-out é uma técnica de validação cruzada amplamente utilizada, popular por sua eficiência e facilidade. Ele separa o conjunto de dados T (de tamanho n) em três subconjuntos mutuamente disjuntos - treinamento T_{tr} , validação T_v e teste T_t de tamanhos n_{tr} , n_v e n_t sucessivamente. Uma vantagem desse método é que a proporção desses três subconjuntos de dados não está estritamente restrita. O modelo é treinado no subconjunto de treinamento T_{tr} , enquanto a T_v do subconjunto de validação é periodicamente utilizada para avaliar o desempenho do modelo durante o treinamento para evitar o excesso de treinamento. O treinamento é interrompido, quando o desempenho na T_v é bom o suficiente ou quando ele pára de melhorar. Quando comparamos mutuamente $m > 1$ modelos computacionais $L_1, \dots, L_m$ uns contra os outros, o subconjunto de teste T_t é usado para obter uma estimativa confiável do desempenho dos modelos. O algoritmo 1 descreve esse processo com mais detalhes. (REITERMANOVÁ, 2010)

Esta abordagem é indicada quando está disponível uma grande quantidade de dados. Caso o conjunto total de dados seja pequeno, o erro calculado na predição pode sofrer muita variação.

2.4.1.2 SVM – Support Vector Machines (Máquinas de Vetores de Suporte)

Os SVMs foram introduzidos por Vapnik no final da década de 1960. Trata-se de uma abordagem baseada na teoria da aprendizagem estatística, um método de classificação geral que, dentre outras aplicações, é utilizado para mineração de banco de dados. (NAGI et al., 2010)

O objetivo principal da metodologia é produzir uma função de decisão ótima (classificador), que determine com precisão a posição de dados novos em uma de duas classes e minimize o erro de classificação usando:

$$f(x) = \text{sgn}(g(x)) \quad (1)$$

Onde: $g(x)$ é o limite de decisão entre as duas classes. Isto é conseguido seguindo o método de minimização do risco estrutural (SRM – Structural Risk Minimization), dado por:

$$R < \frac{t}{n} + \sqrt{\frac{h(\ln(\frac{2N}{h}) + 1) - \ln(\frac{\eta}{4})}{n}} \quad (2)$$

Onde: R é a expectativa de erro de classificação,

t é o número de erros de treinamento,

n é o número de amostras de treinamento e

η é uma medida de confiança.

Quando os dados são separáveis, o primeiro termo em (2) é zero e o segundo termo é minimizado, resultando em um bom desempenho de generalização do SVM. A função $g(x)$ em (1) é o limite de decisão, que é derivado de um conjunto de amostras de treinamento:

$$X = \{x_1, x_2, \dots, x_n\}, x_i \in \mathbb{R}^M, i = 1, 2, \dots, n \quad (3)$$

Onde cada amostra de treinamento x_i possui M características descrevendo uma assinatura particular e pertencente a uma das duas classes.

$$Y = \{y_1, y_2, \dots, y_n\}, y_i \in \{-1, +1\}, i = 1, 2, \dots, n \quad (4)$$

O limite de decisão definido entre duas classes é um hiperplano descrito pela equação:

$$g(x) = (w, x) + b \quad (5)$$

Onde w e b derivam de tal maneira que dados nunca vistos são classificados de maneira correta. Isto é atingido através da maximização da margem de separação entre duas classes. A simplificação desta situação é a transformação em um problema de otimização de programação quadrática (QP). Desta forma surge o parâmetro C , é um parâmetro de regularização que deve ser pré-selecionado para ser o limite de equilíbrio (*tradeoff*) entre os dois termos.

No caso de dados linearmente separáveis, todos os vetores de suporte ficam na margem e, portanto, o número de vetores de suporte (SVs) é menor. Consequentemente, o limite de

decisão $g(x)$ é determinado usando apenas um subconjunto das amostras de treino. Já nos casos em que um limite de decisão linear é inadequado, o SVM mapeia o vetor de entrada para um espaço de característica dimensional maior. Isto é conseguido através da introdução de uma função $K(\cdot, \cdot)$, uma função Kernel. (NAGI et al., 2010)

Os kernels são selecionados com base na estrutura de dados e no tipo de limites entre as classes. Neste trabalho, é utilizada uma função de núcleo baseada na distância euclidiana, o núcleo da função de base radial (RBF):

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2) \quad (6)$$

Esta decisão é tomada considerando que, em geral, o kernel RBF é uma primeira escolha razoável. Esta função eleva os dados a um espaço dimensional maior, permitindo uma classificação com base na relação entre rótulos de classe e atributos não lineares (HSU, CHANG, E LIN, 2003).

Dessa forma, temos apenas como variáveis a definir C e γ , para que seja possível aplicar a metodologia. Onde C , conforme descrito, é o parâmetro que define o erro, enquanto que γ pode ser descrito como o parâmetro que controla a largura da função do kernel RBF e reflete a faixa de distribuição de valores dos dados de treinamento (NAGI et al., 2010).

Não é possível determinar de antemão quais valores de C e γ são melhores para um determinado problema, de tal forma que algum tipo de modelo de seleção (pesquisa de parâmetros) deve ser adotado. O objetivo é identificar bem (C, γ) para que o classificador possa prever com precisão dados desconhecidos (isto é, testes de dados) (HSU, CHANG, E LIN, 2003).

Conforme Nagi et al. (2010), existem muitas abordagens práticas existentes para a seleção de parâmetros (C, γ) , desde definição pelo usuário com base em conhecimento e experiência prévia, otimização assintótica, validação cruzada (CV) e pesquisa em grade (*grid search*).

3 METODOLOGIA

De acordo com Gil (2002), não existe um método que abranja de forma universal todos os tipos de pesquisa científica, entretanto, é possível observar “os elementos considerados essenciais” de forma que sua inclusão na pesquisa, como uma forma de fazer, pode ser considerado um método. Neste projeto, a obra do autor servirá de como roteiro para guiar as espadas da definição da metodologia a ser adotada.

3.1 Delineamento da pesquisa

A pesquisa proposta, se trata de uma abordagem experimental, porém através de um estudo de caso. O objetivo deste método, ainda de acordo com Gil (2002) é investigar hipóteses, descobrir intuições. Neste sentido, o próprio autor aponta que “na maioria dos casos assume a forma de pesquisa bibliográfica ou estudo de caso”.

Pelo fato do problema se tratar de um aspecto específico de uma população ou região a ser determinada, a abordagem julgada como mais adequada é o estudo de campo. Essa decisão encontra apoio na afirmação do autor de que o estudo de campo, ao contrário do levantamento, é mais profundo do que abrangente (GIL, 2002). Assim, é possível se aprofundar no aspecto que não é uma regra geral. Ainda pelo aspecto de análise de comportamento social, tal escolha ainda encontra respaldo em outra afirmação do autor, que afirma que o estudo de caso permite conhecimento profundo e detalhado, tarefa impossível por outros delineamentos.

Dessa forma, o projeto irá desenvolver o estudo de caso com a aplicação de algoritmo baseado em aprendizagem de máquinas, SVM – Support Vector Machines, anteriormente descrita, para detecção de anomalias em matrículas de sistemas de abastecimento de água.

O algoritmo proposto visa classificar as matrículas de determinada localidade entre ligações em que ocorre fraude ou não, baseado nos perfis de consumo das mesmas. O estudo de caso foi realizado com informações de localidade onde todas as ligações analisadas foram pesquisadas fisicamente em busca de fraudes, de forma que os resultados obtidos pelo algoritmo possam ser comparados com as informações encontradas em campo.

3.2 Aquisição de dados

Para alimentar o algoritmo SVM foram utilizadas as séries históricas de consumo de quatro DMCs – Distritos de Medição e Controle pertencentes a quatro regiões distintas do sistema de abastecimento de água de uma localidade de Mato Grosso do Sul. Todas essas áreas são abastecidas por um ponto individual da rede de distribuição e possuem macromedição. Suas matrículas são cadastradas no sistema comercial de forma a tornar possível a comparação entre o volume macromedido e micromedido, e assim, obter as perdas de água de cada uma delas individualmente.

Tabela 01 – Dados das áreas de estudo.

DMC	Ligações	Extensão de rede	Índice de Fraudes
DMC-056	1.943	45,420 km	1,49%
DMC-057	1.486	28,447 km	1,75%
DMC-059	1.392	15,899 km	4,16%
DMC-060	3.893	49,263 km	2,00%

Os dados históricos das matrículas analisadas foram obtidos/gerados através do sistema de informações comerciais da Companhia de Saneamento. Os dados obtidos representam o histórico de 8.714 matrículas durante um período de 60 meses, ou seja, de abril de 2011 a abril de 2016. As informações obtidas para o desenvolvimento do presente trabalho incluíram informações de ordens de serviço abertas por matrícula, durante todo o período de pesquisa, consumo mensal medido, notificações por fraude referente às matrículas, tipo de ligação (residencial, comercial, industrial), endereço, coordenadas geográficas das matrículas.

Além disso, mediante pesquisa realizada entre julho de 2016 a dezembro, foi realizada pesquisa completa em campo de todas as matrículas analisadas, identificando os clientes com ligação fraudada e fornecida lista das matrículas fraudadas.

Os dados foram fornecidos em formato de planilha eletrônica e continham 193 matrículas notificadas por fraude neste período, ou seja, 2,21% do total.

3.3 Seleção e extração dos dados

Os recursos foram selecionados a partir dos dados do sistema comercial pré-processados para construir o classificador SVM. Além das séries temporais de dados de consumo mensal por matrícula, também foram escolhidas outras características de cadastro ou geradas pela

companhia ao longo do período de estudo, convertidas em um arquivo no formato “.csv” (comma separated values). A coluna indicando se em determinada matrícula foi encontrada fraude na pesquisa em campo foi gerada de forma binária, representando matrículas com fraude (1) e sem fraude (0).

As demais informações escolhidas como relevantes foram as seguintes:

a) Geolocalização

Para cada matrícula, obteve-se as coordenadas geográficas métricas UTM (Universal Transversa de Mercator) que descrevem a posição espacial (X,Y) de cada uma, visando caracterizar fraudadores por regiões através de análises mais detalhadas. As coordenadas foram extraídas do software SIG (Sistema de Informações Geográficas) da companhia e as análises foram desempenhadas plotando as coordenadas em um gráfico através do software *Tableau Public 10.5* representando a distribuição geográfica, verificando se as ocorrências de fraude estavam agrupadas em determinadas regiões.

b) Tipos de economia

Esta característica descreve se a matrícula está relacionada a um Comércio, Indústria ou Residência. Esta informação é obtida diretamente da base de dados.

c) Média de consumos

Para cada matrícula, obteve-se a média do consumo de todos os meses medidos. Esta característica teve seus resultados normalizados para valores entre 0 e 1, para melhorar a performance do modelo.

d) Frequência de zeros

Como apenas obtendo a média de consumo de cada matrícula podemos perder muita informação, é importante utilizar outras características relacionadas ao consumo que caracterizem de alguma forma fraudes. Esta característica mede a quantidade de consumos zero no histórico da matrícula.

e) Ocorrência

Ocorrência de maior frequência da matrícula.

f) Endereço

Além das coordenadas geográficas, utilizou-se as informações de nome e número da rua associados a matrícula.

3.4 Normalização de dados

Os dados das características selecionadas precisam ser representados em uma escala normalizada para capacitação e validação SVM. Portanto, todos os registros foram normalizados usando:

$$NL = \frac{L - \min(L)}{\max(L) - \min(L)} \quad (7)$$

Onde: L representa o consumo pontual de um determinado cliente em uma determinada série de consumo e

$\min(L)$ e $\max(L)$ representam o consumo mínimo e máximo no perfil de consumo do mesmo.

3.5 SVM Classificador e otimização

Para aplicação do algoritmo SVM foi utilizado o software de programação de cálculos e cálculos estatísticos *RStudio 1.1.414*. Neste software utilizou-se a interface SVM do pacote *e1071*, que emula a aplicação LIBSVM desenvolvida por Chih-Chung Chang and Chih-Jen Lin.

Como a relação entre as duas classes é desequilibrada (Classe positiva com 193 amostras e Classe negativa com 8.521 amostras), portanto, foi necessário lançar mão de um método para otimizar o classificador SVM, equilibrando os casos listados para identificar os perfis de consumo na proporção da amostra.

Considerando que há muito mais não-fraudes do que fraudes, isso normalmente pode prejudicar o aprendizado do classificador. O método utilizado para evitar este problema foi a redução da amostra (*subsample*) dos exemplos da classe negativa (não-fraudes) como forma de balancear o conjunto de dados e melhorar a performance do classificador.

A modificação da amostragem é feita aleatoriamente no conjunto da classe negativa em diversas porcentagens: 50%, 30%, 20%, 10% e 5%. O ponto positivo é que torna possível

balancear o conjunto de dados de forma mais adequada, melhorando teoricamente a desempenho do classificador. Em contrapartida, corre-se o risco de estar perdendo informação da classe negativa, dado que estamos diminuindo o conjunto de observação apresentado ao algoritmo.

A precisão de classificação ótima do classificador C-SVM foi estimada pela otimização do parâmetro do kernel RBF, γ e do parâmetro de penalidade de erro, C . Para este estudo, o método de *grid search* foi utilizado para encontrar estes valores. No método de *grid search*, as sequências de parâmetros exponencialmente crescentes (C , γ) foram usadas para identificar os parâmetros C-SVM obtendo a melhor precisão por validação cruzada.

Neste método vários pares de valores (C , γ) são testados e aquele com a melhor precisão de validação cruzada é escolhido. O tempo computacional necessário para encontrar bons parâmetros pelo método *grid search* não é muito maior que para outros métodos avançados, uma vez que existem apenas dois parâmetros. Além disso, a busca em grade pode ser facilmente paralelizada porque cada (C , γ) é independente (HSU, CHANG, E LIN, 2003). O método *grid search* foi executado utilizando a função *tune.svm* do pacote de exemplos e1071 do RStudio.

Por fim, para validar a eficiência do classificador SVM, sua precisão (acurácia) revocação (*recall*) são calculados usando as seguintes expressões:

$$Precisão = \frac{T_p}{T_p + F_p} \times 100\% \quad (8)$$

Onde: T_p representa o número de amostras corretamente classificadas (verdadeiros positivos) pelo SVM e

F_p representa o número de amostras classificadas equivocadamente (falsos positivos).

A precisão é, intuitivamente, a capacidade do classificador de não rotular como positivo uma amostra que é negativa, enquanto a revocação (*recall*) ou sensibilidade de detecção do classificador é definido pela seguinte expressão:

$$Recall = \frac{T_p}{T_p + F_n} \times 100\% \quad (9)$$

Onde: T_p é definido como o número de verdadeiros positivos;

F_n representa o número de amostras classificadas equivocadamente (falsos negativos).

Esta métrica mede a capacidade do classificador de encontrar todas as amostras positivas. Neste caso, ela mede a porcentagem de matrículas fraudadas que foram encontradas.

É importante encontrar um equilíbrio entre os dois, pois um sistema com revocação elevada, mas baixa precisão retorna muitos resultados, mas a maioria dos seus rótulos previstos são incorretos quando comparados aos rótulos de treinamento. Um sistema com alta precisão, mas baixa revocação é exatamente o oposto, retornando poucos resultados, mas a maioria dos seus rótulos previstos são corretos quando comparados aos rótulos de treinamento. Um sistema ideal com alta precisão e alto índice de revocação retornará muitos resultados, com todos os resultados rotulados corretamente.

4 RESULTADOS E DISCUSSÕES

Com base na metodologia proposta para o desenvolvimento do algoritmo classificador SVM, considerando os dados disponíveis e os testes de validação aplicados relatados anteriormente, foram obtidos os seguintes resultados quanto as informações possíveis de interpretação e a qualidade da predição do classificador SVM aplicado:

4.1 Características

Em resumo, foram desempenhadas algumas análises acerca dos conjuntos de dados com base nas características utilizadas como dados de entrada do algoritmo, com o intuito de responder algumas questões, como qual a distribuição das classes no conjunto de dados ou se as fraudes costumam aparecer em lugares comuns. Os resultados obtidos através desta análise foram os descritos abaixo.

4.1.1 Geolocalização

Através da análise da distribuição espacial das matrículas identificadas com fraudes, observa-se em representação gráfica que a característica escolhida é pouco conclusiva para os 04 setores estudados. Em alguns casos observa-se pequena concentração de fraudes em regiões esparsas, porém não o suficiente para concluir-se que há uma relação direta entre as fraudes e as regiões em que foram identificadas.

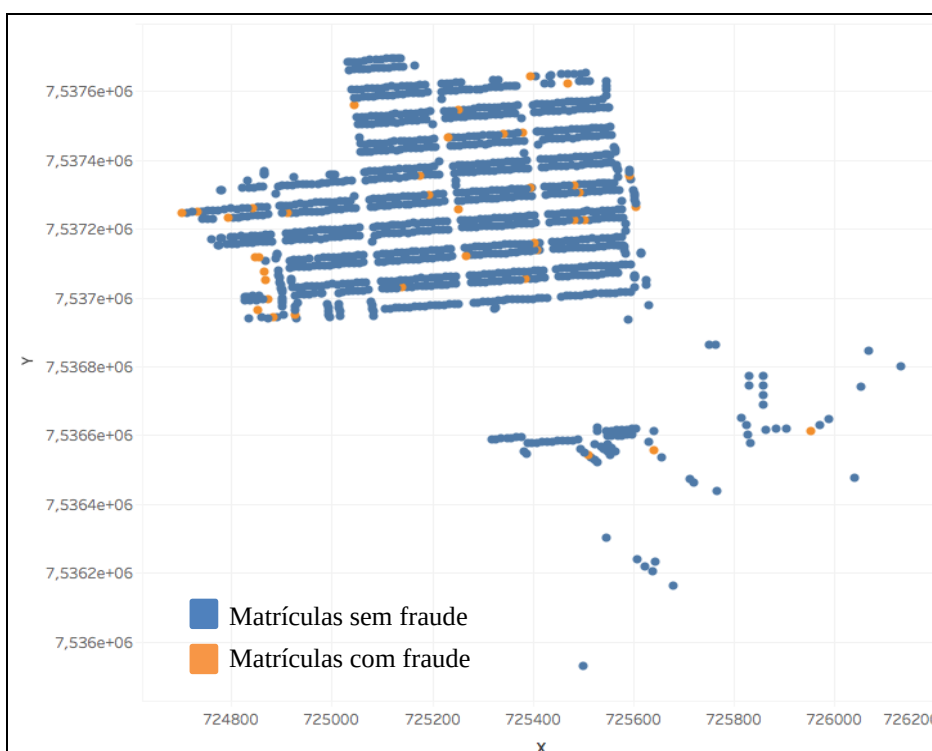


Figura 03 – Distribuição espacial das fraudes no DMC 056.

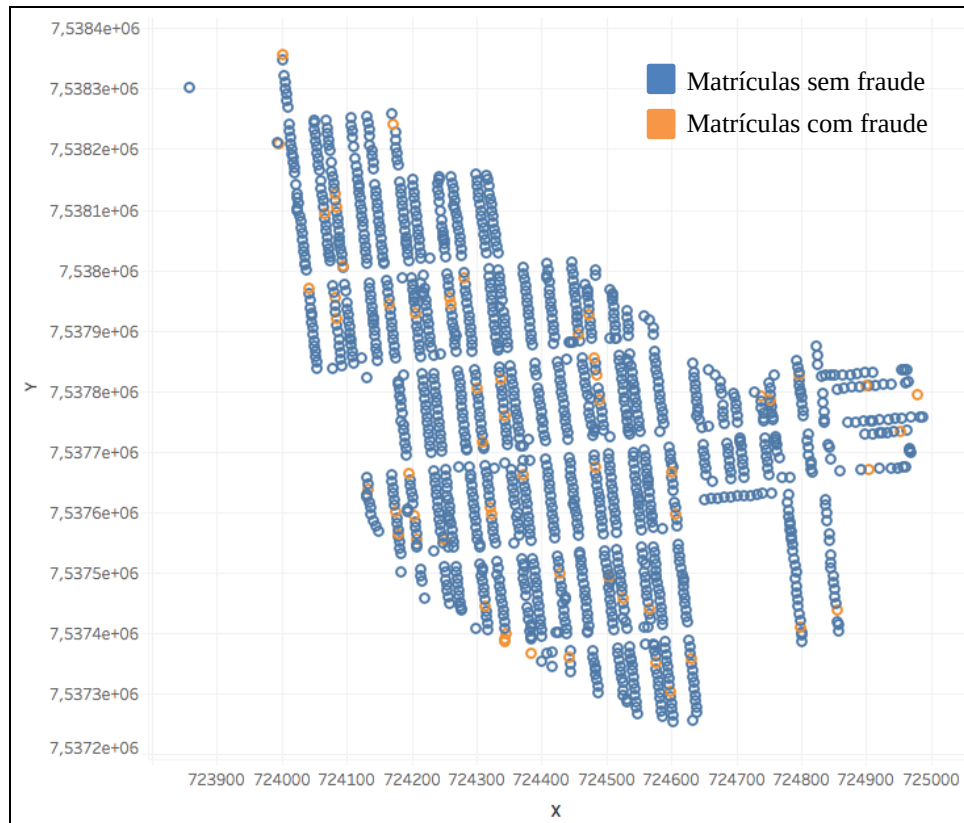


Figura 04 – Distribuição espacial das fraudes no DMC 057.



Figura 05 – Distribuição espacial das fraudes no DMC 059

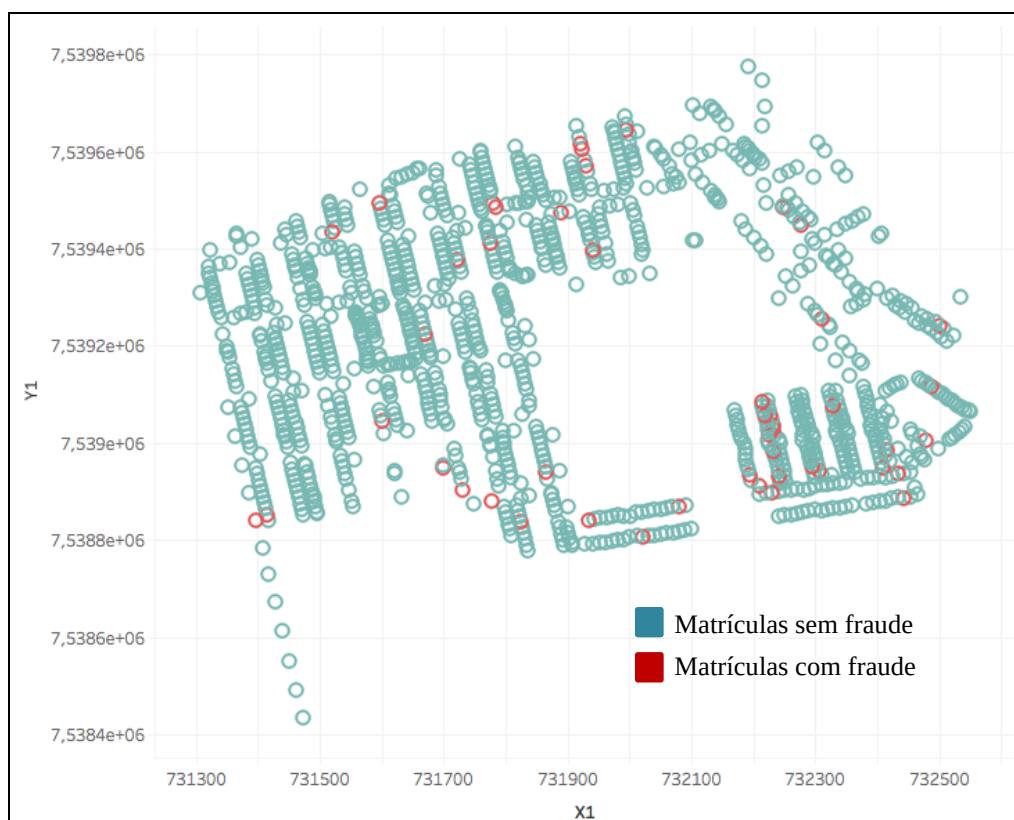


Figura 06 – Distribuição espacial das fraudes no DMC 060

4.1.2 Médias de Consumo por Classes

Ao contrário da característica de localização, quando observamos os gráficos de distribuição do consumo médio separado por classes de investigação (ligações não fraudadas e ligações fraudadas), passa a ser mais evidente um comportamento chave para que o algoritmo identifique e classifique determinada matrícula entre um possível positivo ou negativo para fraudes.

Com exceção do DMC-057, existe uma diferença perceptível de distribuição das médias de consumo das classes positivas (fraudadas) em relação aos seus vizinhos da classe negativa, representando uma dimensão importante a ser considerada pelo algoritmo classificador.

características, objetivando validar esses procedimentos e aproveitar tal conhecimento como dimensão determinante na classificação matrículas analisadas.

Observa-se através dos gráficos que, de fato, a grande incidência de consumos zeros ou demais ocorrências comerciais está diretamente relacionada às ligações fraudadas (classe positiva). Nos gráficos a seguir, o eixo vertical indica a incidência destas ocorrências comerciais em relação às matrículas elencadas na classe positiva, enquanto a intensidade da cor azul representa em quais destas ocorrências se observou maior repetição de consumo zero.

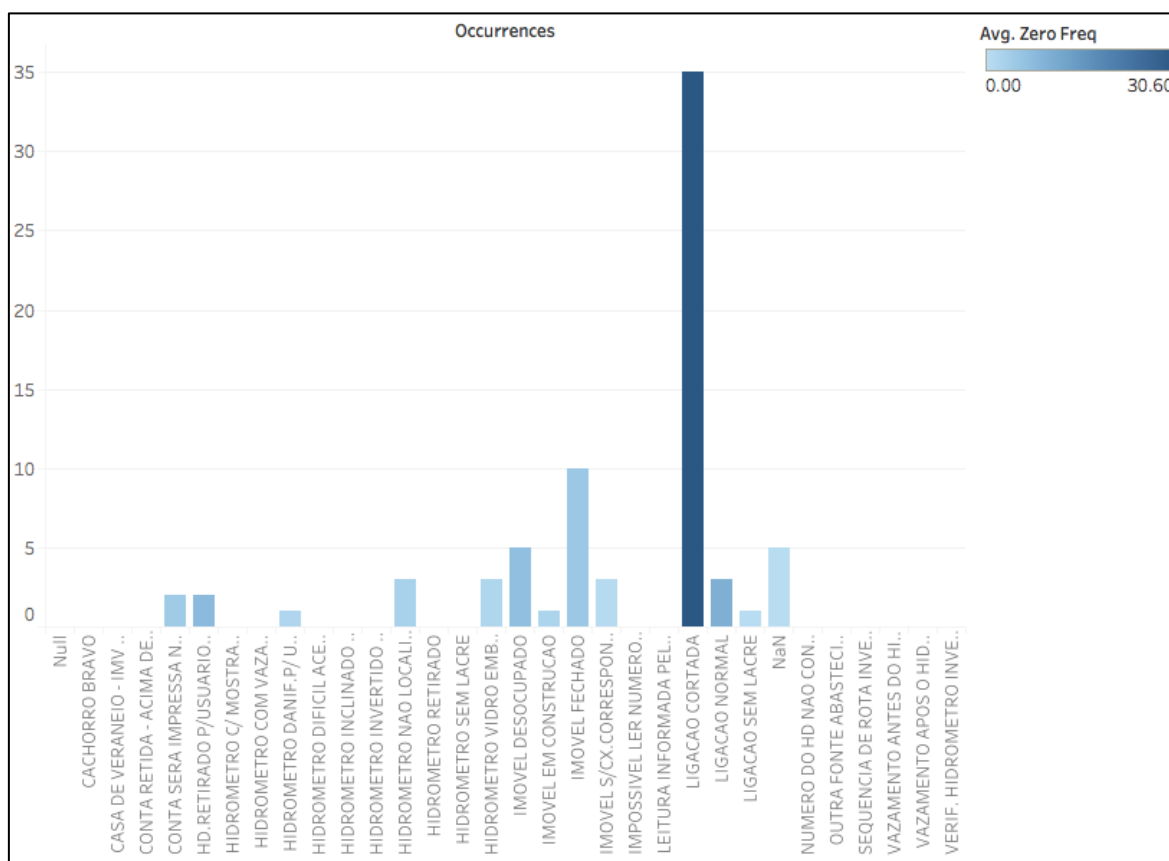


Figura 08 – Ocorrências comerciais x Fraudes x Consumo zero – DMC-056

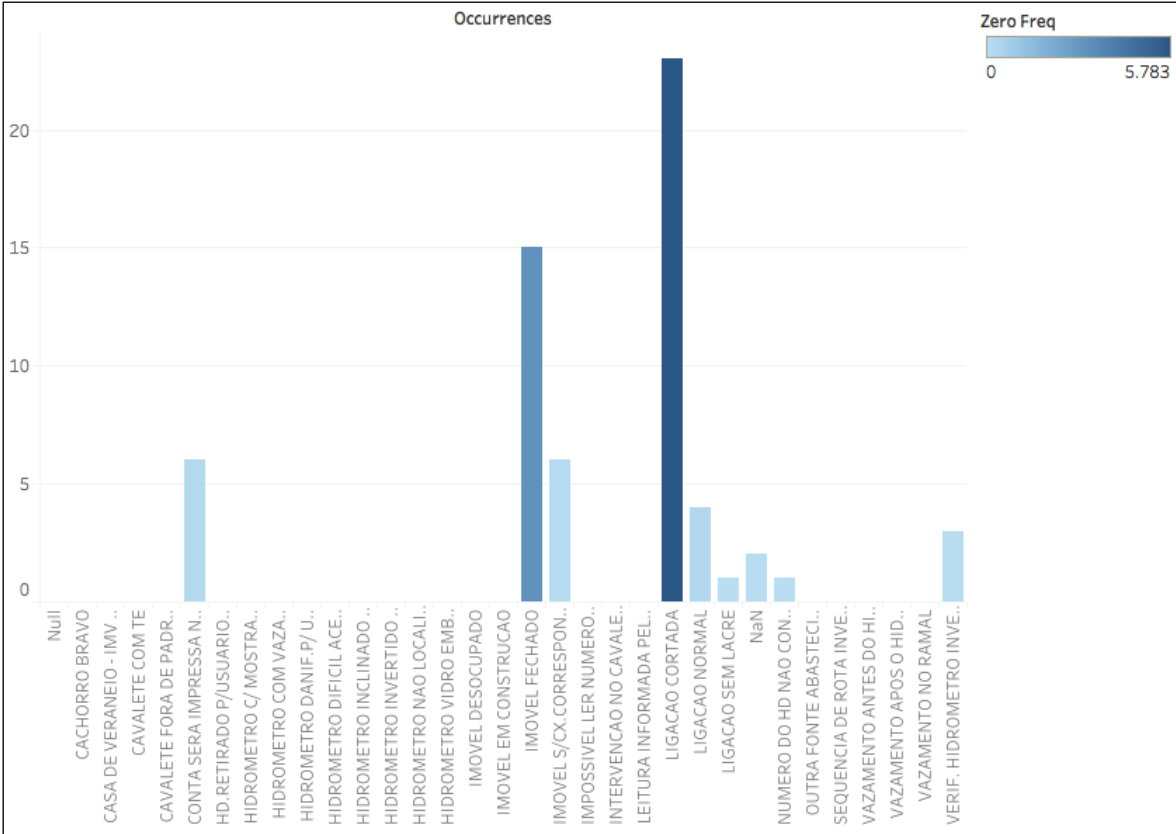


Figura 09 – Ocorrências comerciais x Fraudes x Consumo zero – DMC-057

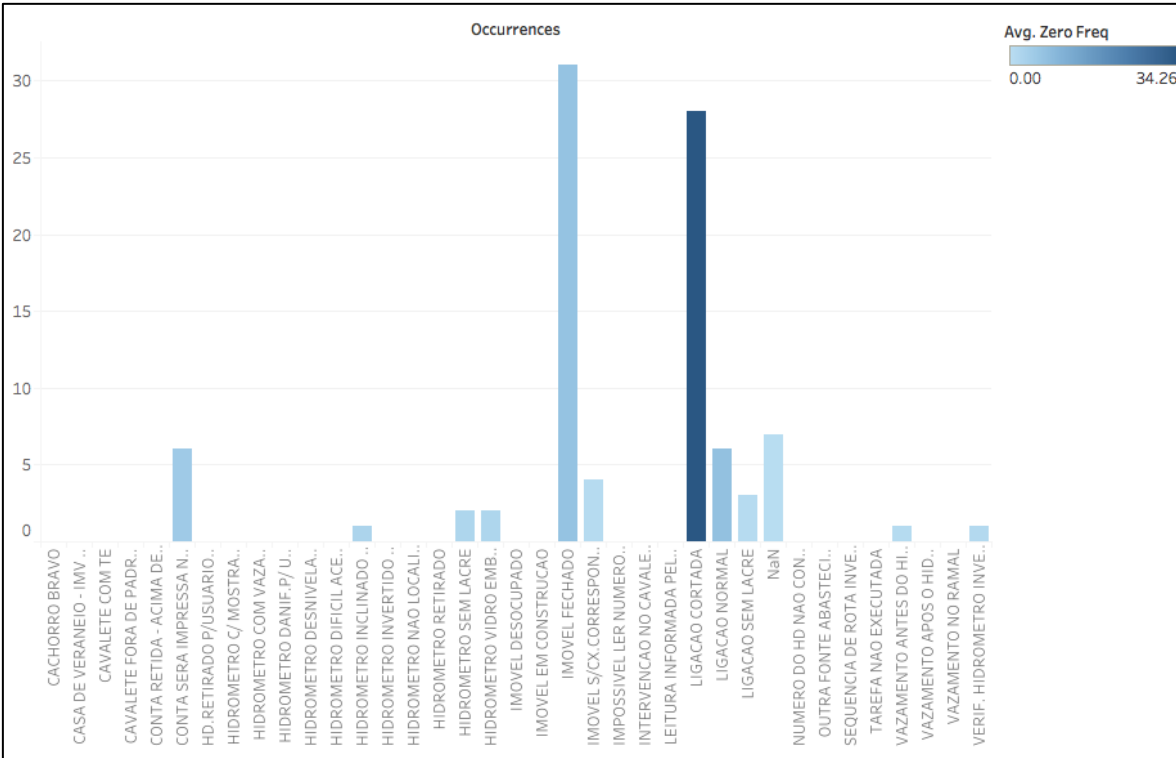


Figura 10 – Ocorrências comerciais x Fraudes x Consumo zero – DMC-059

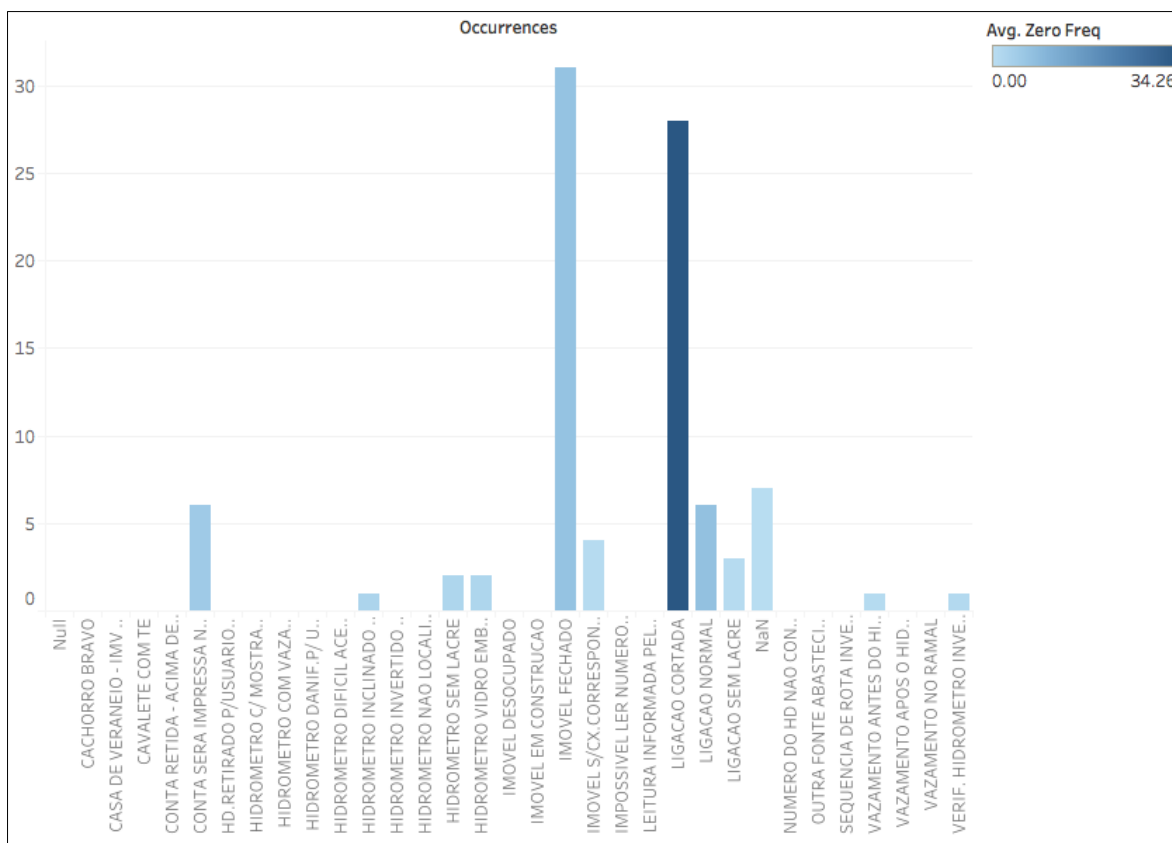


Figura 11 – Ocorrências comerciais x Fraudes x Consumo zero – DMC-060

Observa-se que a maioria das ocorrências de consumo zero foi observado em situações de ligação cortada, entretanto, grande parte das ligações fraudadas possuíam registros comerciais de “imóvel fechado” e “ligação cortada”.

4.2 Verificação de Desempenho do Algoritmo

Após a aplicação do algoritmo com base nas características elencadas, realizou-se avaliação da eficiência do mesmo sob a ótica dos parâmetros de precisão e revocação (*recall*), conforme demonstrado a seguir.

4.2.1 Precisão

A tabela 02 demonstra os resultados obtidos para o classificador SVM a partir do uso de todas as instâncias do conjunto de dados (sub-0, ou sem uso de subamostragem), e também com diversas situações de subamostragem, sendo 50% dos dados da classe negativa (sub-50), 10% da classe negativa (sub-10), 5% da classe negativa (sub-5) e a mesma quantidade de dados de classe negativa e positiva (sub-50/50).

Tabela 02 – Resultados de precisão para os grupos de dados completos e após subamostragem.

	DMC 056	DMC 057	DMC 059	DMC 060
Sub-0	0,0735	0,0588	0,0891	0,0462
Sub-50	0,1522	0,1087	0,1818	0,2857
Sub-10	0,3000	0,4167	0,5200	0,5385
Sub-5	0,7059	0,4615	0,5313	0,6000
Sub-50/50	0,6667	0,4375	0,7368	0,6429

A seguir, os dados de resultados de precisão para a utilização de todos os dados e por arranjos de subamostragem são demonstrados graficamente visando melhor compreensão das situações delimitadas.

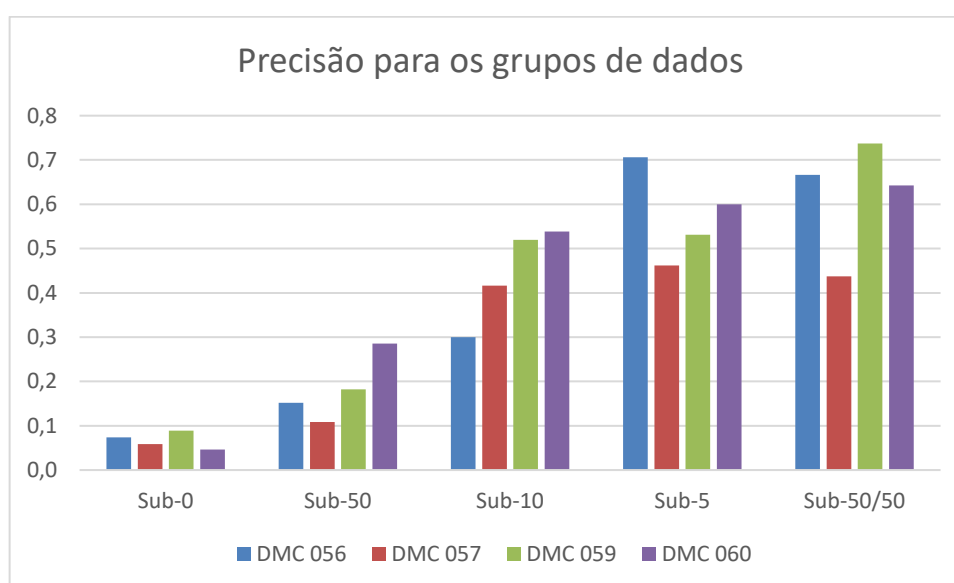


Figura 12 – Precisão obtida através dos grupos de dados propostos.

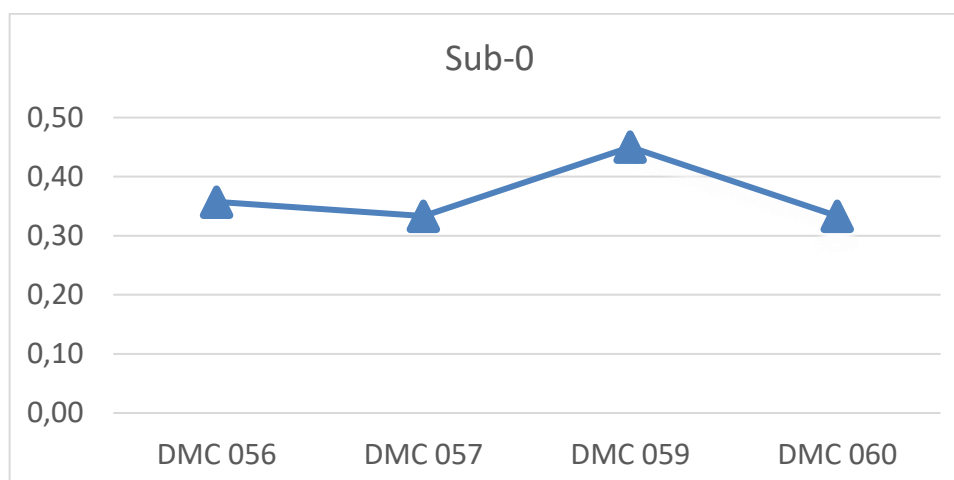


Figura 13 – Precisão obtida sem o uso de subamostragem.

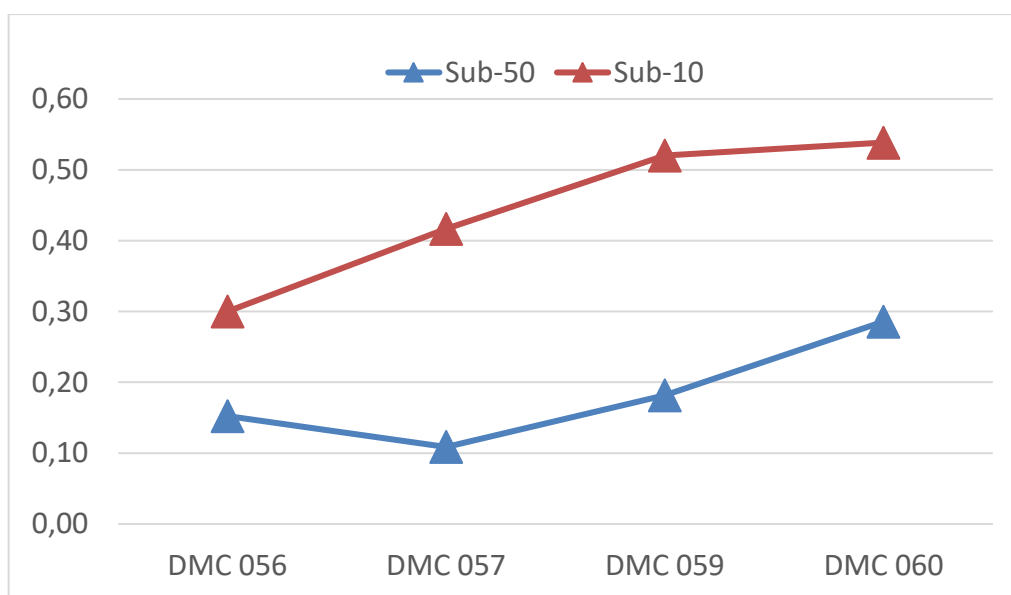


Figura 14 – Precisão obtida com o uso de subamostragem (Sub-10 e Sub-50).

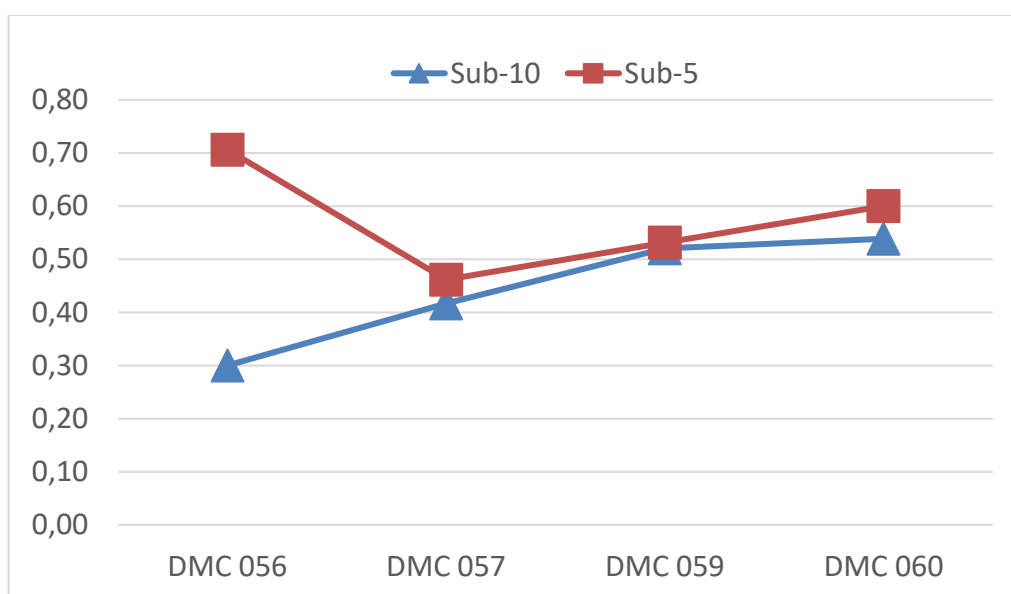


Figura 15 – Precisão obtida com o uso de subamostragem (Sub-10 e Sub-5).

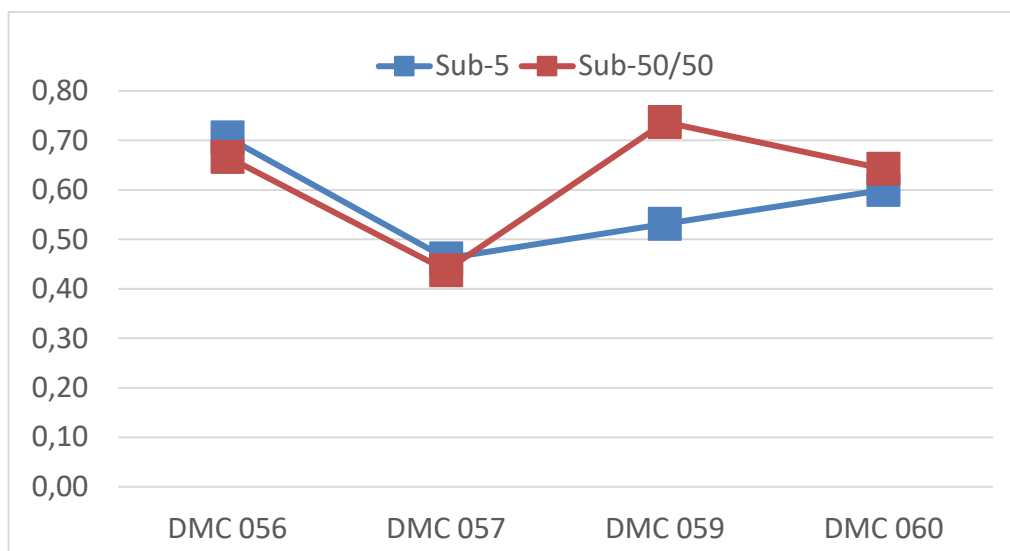


Figura 16 – Precisão obtida com o uso de subamostragem (Sub-5 e Sub-50/50).

Fica evidente observando os gráficos apresentados que, devido ao desequilíbrio entre as classes positivas e negativas (fraudes e não fraudes), com o uso do artifício da subamostragem temos melhora significativa na precisão do algoritmo, partindo de uma média de 6,7% de precisão para 62,1% no caso do subamostragem de 50/50.

Observa-se também que, na maioria das possibilidades de proporções de dados, a precisão do DMC-057 se apresentou abaixo das demais.

4.2.2 Revocação

A tabela 03 demonstra os resultados de revocação obtidos para o classificador SVM a partir do uso de todas as instâncias do conjunto de dados (sub-0, ou sem uso de subamostragem), e também com diversas situações de subamostragem, sendo 50% dos dados da classe negativa (sub-50), 10% da classe negativa (sub-10), 5% da classe negativa (sub-5) e a mesma quantidade de dados de classe negativa e positiva (sub-50/50).

Tabela 03 – Resultados de revocação para os grupos de dados completos e após subamostragem.

	DMC 056	DMC 057	DMC 059	DMC 060
Sub-0	0,3571	0,3333	0,4500	0,3333
Sub-50	0,5833	0,4167	0,3529	0,5882
Sub-10	0,6429	0,5833	0,6190	0,6364
Sub-5	0,6316	0,5000	0,9444	0,8182
Sub-50/50	0,8333	0,6364	0,7281	0,9000

A seguir, os dados de resultados de revocação para a utilização de todos os dados e por arranjos de subamostragem são demonstrados graficamente visando melhor compreensão das situações delimitadas.

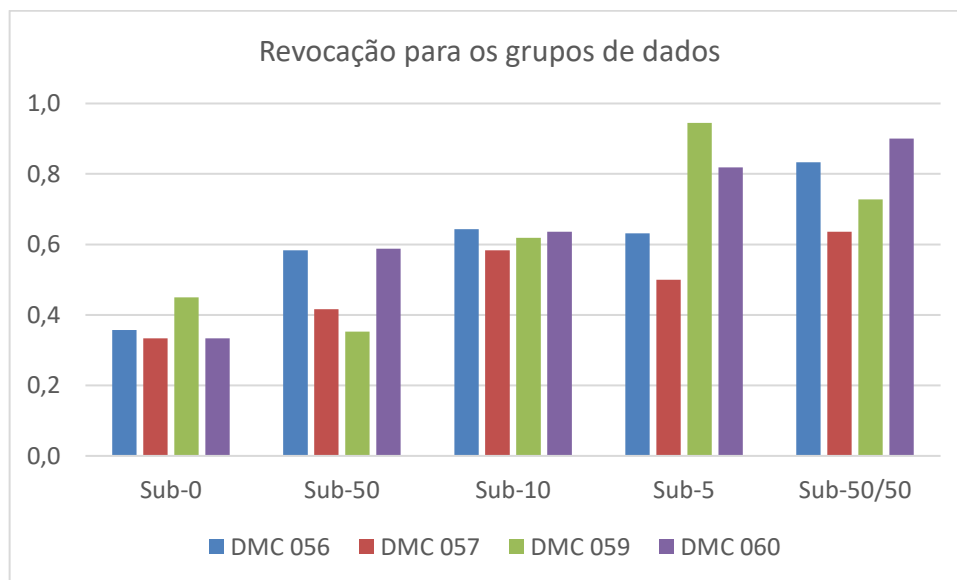


Figura 17 – Revocações obtidas dos grupos de dados propostos.

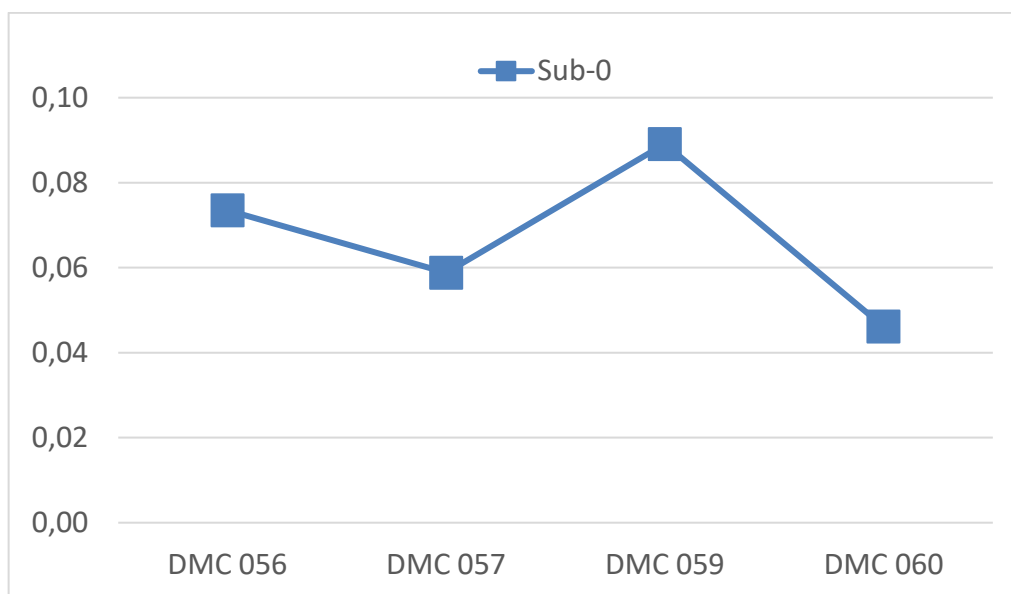


Figura 18 – Revocação obtida sem o uso de subamostragem.

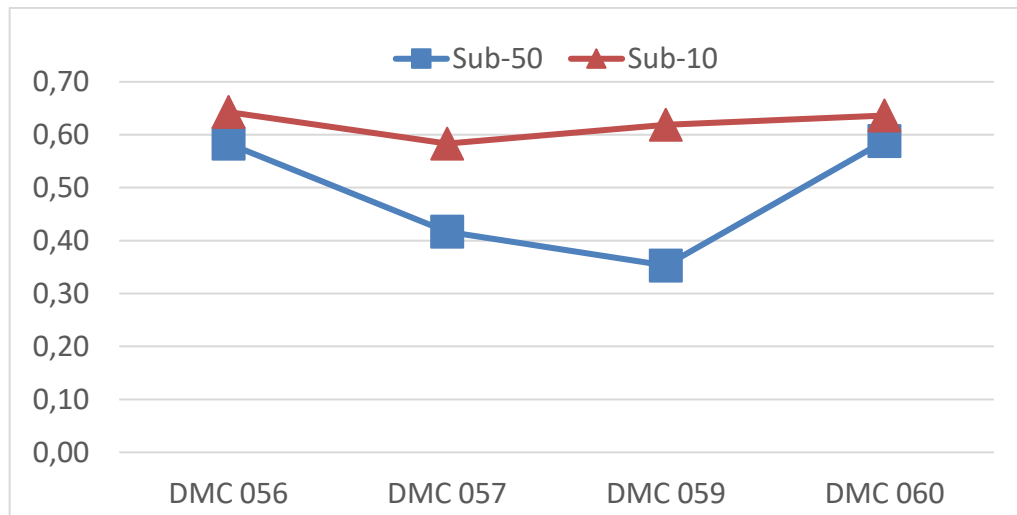


Figura 19 – Revocação obtida com o uso de subamostragem (Sub-10 e Sub-50).

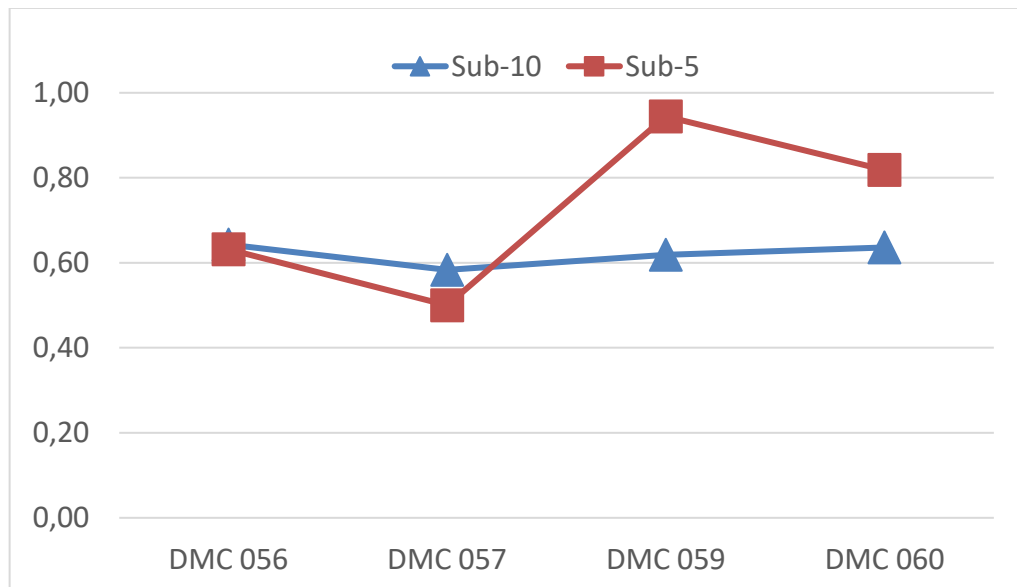


Figura 20 – Revocação obtida com o uso de subamostragem (Sub-10 e Sub-5).

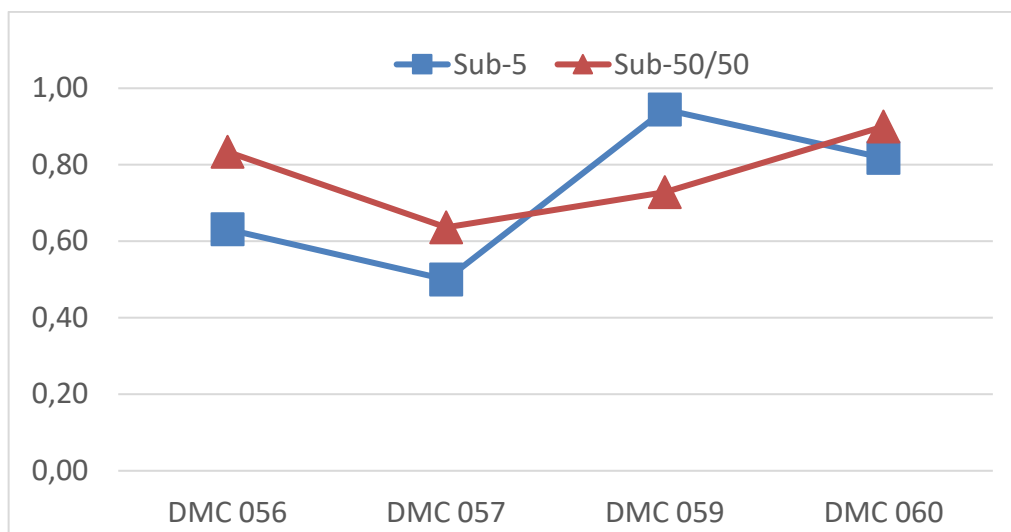


Figura 21 – Revocação obtida com o uso de subamostragem (Sub-5 e Sub-50/50).

Da mesma forma como para a precisão, também se evidencia através dos gráficos apresentados que, devido ao desequilíbrio entre as classes positivas e negativas (fraudes e não fraudes), com o uso do artifício da subamostragem obtém-se melhora significativa no índice de revocação do algoritmo, partindo de uma média de 36,8% de precisão para 77,4% no caso do subamostragem de 50/50.

Observa-se também que, na maioria das possibilidades de proporções de dados, a precisão do DMC-057 se apresentou abaixo das demais.

5 CONCLUSÃO

Com base na hipótese levantada inicialmente, de que é possível traçar, através de séries temporais, ou seja, dados históricos de consumo das ligações de uma determinada localidade, um perfil que seja comum à parcela de consumidores que praticam o consumo não autorizado, e assim criar mecanismos automatizados de identificação de ligações fraudadas para combater estas práticas, foi realizada com sucesso a revisão bibliográfica sobre o tema e a avaliação da aplicabilidade do referencial teórico apresentado.

Neste trabalho apresentou-se uma nova abordagem para a detecção de fraudes em ligações de sistemas de abastecimento de água usando uma técnica baseada em inteligência artificial, a partir de uma série de dados comumente registrados nos sistemas de informação comercial das companhias de saneamento, como histórico de consumo, informações de fraude (adulteração de medidor, bypass de contador, etc.) entre outras atividades. O presente estudo aplicou a técnica de classificação de padrões através de aprendizagem de máquina, algoritmo de máquina de vetores de suporte - SVM, para detectar e identificar padrões de consumo de clientes com fraude.

Os resultados experimentais obtidos indicam que o algoritmo classificador com base em SVM proposto pode ser usado para detecção confiável e eficiente de anormalidades e atividades de fraude dentro de serviços de fornecimento de água.

O método proposto para detecção de clientes de fraude se mostrou promissor, em primeiro lugar, pois o SVM pode considerar diversas características e aplica-las em um plano n-dimensional, o que lhe confere alta capacidade de discriminação. Em segundo lugar, os resultados obtidos indicaram alta eficiência na classificação dos dados de teste e validação.

Em terceiro lugar, e talvez o argumento mais forte a favor da metodologia desempenhada, é considerar-se que, com o método de pesquisa tradicional adotado pelas companhias de saneamento, em que todas as ligações são visitadas em campo implicando em grande necessidade de mão de obra e custos, apresentou um índice de eficiência muito inferior ao algoritmo SVM.

Enquanto a companhia em questão levou em torno de sete meses para visitar as 8.714 matrículas com detecção de 2,21% de matrículas com fraude, o algoritmo SVM, em questão de horas, é capaz de gerar uma relação de matrículas a serem verificadas com precisão média de até 62% e um índice de revocação média de até 77,4%.

Conclui-se que há a necessidade de emprego da técnica de subamostragem para eliminar o desequilíbrio entre as classes, e assim, garantir os resultados de eficiência identificados. Quanto as proporções de subamostragem verificadas, é possível afirmar que, para o caso estudado, a melhor situação verificada é para uma proporção de 5% do total da amostra de instâncias negativas. Mesmo considerando que a proporção de 50% de dados negativos para 50% de dados positivos tenha apresentado uma média de valores de revocação e precisão mais elevados em relação a sub-5, observa-se que com o uso de 5% de dados negativos temos resultados máximos para todos os DMCs, com exceção do DMC 057, que apresentou discrepância e baixa eficiência em relação aos demais para quase todos os resultados obtidos.

Além disso, a escolha da proporção de subamostragem de 5% dos dados negativos visa obter valores de revocação maiores, já que é preferível visitar mais pontos e encontrar alguns falsos positivos do que visitar menos matrículas em decorrência de falsos negativos.

Por fim, com base nos resultados obtidos, concluímos este trabalho propondo futuras análises com o objetivo de refinar a metodologia desenvolvida:

- Investigar os fatores que influenciaram a baixa performance dos resultados obtidos para o DMC 057;

- Incluir novas características presentes no sistema de cadastro e informações comerciais (presença de fonte alternativa de abastecimento e cobrança de serviços de esgoto, por exemplo), de forma a refinar os resultados obtidos neste trabalho.

Assim, o presente trabalho demonstrou a relevância para do tema estudado para o mercado de saneamento, para que possam tornar mais eficiente a identificação de perdas aparentes em ligações dos sistemas que operam, reduzindo o mau uso dos recursos necessários para a produção de água tratada, dentre eles a energia, tornando-as mais eficientes do ponto de vista econômico, social e ambiental.

REFERÊNCIAS

ABES – Associação Brasileira de Engenharia Sanitária e Ambiental; **Controle e redução de perdas nos sistemas públicos de abastecimento de água: posicionamento e contribuições técnicas da ABES**, 2015. 99 p. Disponível em < http://abes-dn.org.br/pdf/28Cbesa/Perdas_Abes.pdf>. Acesso em: 06 outubro. 2016.

ABNT- Associação Brasileira de Normas Técnicas. **NBR ISO 24.510/2012 - Atividades relacionadas aos serviços de água potável e de esgoto — Diretrizes para a avaliação e para a melhoria dos serviços prestados aos usuários**. Rio de Janeiro, ABNT, 2012.

ALLIANCE TO SAVE ENERGY. **Água e Energia: Aproveitando as oportunidades de efficientização de água e energia não exploradas nos sistemas de água municipais**. Washington, USA; 2002. Disponível em < http://www.lenhs.ct.ufpb.br/html/downloads/livros/agua_energia/aguaeenergia.pdf>. Acesso Em: 23 mai. 2015.

ANDRADE FILHO, L. S. BOMBAS E ESTAÇÕES ELEVATÓRIAS. IN: HEBER PIMENTEL GOMES. (ORG.). **Sistemas de Bombeamento: Eficiência Energética**. 1Ed. João Pessoa: Editora Universitária - UFPB, 2009, V. 1, P. 45-112.

AUBUCHON, Craig P.; ROBERSON, J. A.; **Evaluating the embedded energy in real water loss**. American Water Works Association, mar. 2014, Vol.106:3, pp.129-138.

BAKIRLI, G.; BIRANT, D.; MUTLU, E.; KUT, A.; DENKTAŞ, L.; ÇETIN, D.; **Mineração de Dados Solutions for Local Municipalities**. Electronic Journal of e-Government. 2012. Vol. 10:2. Dez. 2012. pp.97-106.

BRASIL. **Lei nº 13.329, de 1º de agosto de 2016. Altera a Lei no 11.445, de 5 de janeiro de 2007**, que estabelece diretrizes nacionais para o saneamento básico, para criar o Regime Especial de Incentivos para o Desenvolvimento do Saneamento Básico - REISB, com o objetivo de estimular a pessoa jurídica prestadora de serviços públicos de saneamento básico a aumentar

seu volume de investimentos, por meio da concessão de créditos relativos à contribuição para o Programa de Integração Social - PIS e para o Programa de Formação do Patrimônio do Servidor Público - PASEP e à Contribuição para Financiamento da Seguridade Social - COFINS. Brasília, 2016. Disponível em: < http://www.planalto.gov.br/CCIVIL_03/_Ato2015-2018/2016/Lei/L13329.htm > . Acesso em: 25 out. 2016

BRASIL. MINISTÉRIO DAS CIDADES. SECRETARIA NACIONAL DE SANEAMENTO AMBIENTAL – SNSA. SNIS - SISTEMA NACIONAL DE INFORMAÇÕES SOBRE SANEAMENTO: **20º Diagnóstico dos Serviços de Água e Esgotos – 2013**. Brasília: SNSA/MCIDADES, 2015. 212 p. Disponível em <<http://www.snis.gov.br>>. Acesso Em: 14 Maio. 2016.

BRASIL MINISTÉRIO DE MINAS E ENERGIA. **Anuário Estatístico de Energia Elétrica 2014**. Rio de Janeiro: EPE, 2014. 212 p.

BRASIL MINISTÉRIO DE MINAS E ENERGIA. **Anuário Estatístico de Energia Elétrica 2015 – Ano Base 2014**. Rio de Janeiro: EPE, 2015. 232 p.

CARDELL-OLIVER, Rachel. **Water use signature patterns for analyzing household consumption using medium resolution meter data**. Water Resources Research, Dez. 2013, Vol. 49, pp. 8589–8599.

CARDELL-OLIVER, R.; WANG, J.; GIGNEY, H.; **Smart Meter Analytics to Pinpoint Opportunities for Reducing Household Water Use**. Journal of Water Resources Planning and Management. Jun. 2016. Vol. 142:6. p. 9.

CARDOSO. A. P. P. M.; **Redução de perdas reais em sistemas de abastecimento de água através do controlo avançado de pressão – aplicação a um subsistema da cidade do Porto**. Dissertação (Mestrado em Engenharia do Ambiente) – Faculdade de Engenharia, Universidade do Porto, Porto. 2013. Disponível em: < <https://repositorio-aberto.up.pt/bitstream/10216/69239/2/43972.pdf> >. Acesso em: 07 de dez. 2016.

CNI - CONFEDERAÇÃO NACIONAL DA INDÚSTRIA. **Perfil da Indústria nos Estados 2014.** – Ed. Rev. – Brasília: CNI, 2014. 214 p.

COLOMBO, A ; KARNEY, B. **Energy and costs of leaky pipes Toward comprehensive Picture.** Journal of Water Resources Planning and Management, Nov. 2002, Vol.128. pp.441-450

FETTERMANN, D. C.; GUERRA, K. C.; MANO, A. P.; MARODIN, G. DE A.; **Uma sistemática para detecção de fraudes em empresas de abastecimento de água.** Interciencia. Fev. 2015. Vol. 40:2, pp. 114-120.

GIL, Antonio Carlos. **Como elaborar projetos de Pesquisa.** 4. ed. São Paulo: Atlas, 2002..

GOMES, H. P., DE CARVALHO, P. S. O. (ORGANIZADORES). IN: GOMES, H. P. (ORG.). **Sistemas de Bombeamento: Eficiência Energética.** João Pessoa: Editora UFPB, 2009. 460p.

GONÇALVES, R. F. **Conservação de Água e Energia em Sistemas Prediais Públicos de Abastecimento de Água.** Rio de Janeiro. 2009. 352 p.

GOULAO M.; BRITO E ABREU F., **Cross-validation of Metrics for Software Components.** IEEE Latin America Transactions, mar. 2005, vol. 3, no. 1, pp. 127-136.

GREENPEACE. **[R]Evolução Energética: A Caminho do Desenvolvimento Limpo.** São Paulo: Greenpeace, 2010. Disponível Em: <<http://www.greenpeace.org/brasil/global/brasil/report/2010/11/revolucaoenergeticadeslimpo.pdf>>. acesso em: 25 maio. 2015.

HENRIQUES, H.O., BARBERO, A.P.L., RIBEIRO, R.M., FORTES, M.Z., ZANCO, W., XAVIER, O.S., AMORIM, R.M. **Development of adapted ammeter for fraud detection in low-voltage installations.** Measurement. Volume 56. Out, 2014, pp. 1–7.

HIGHTOWER, M.; **ENERGY MEETS WATER: Two basic needs have a common cause: New Technologies to assure a sustainable future.** The American Society of Mechanical Engineers. Jul. 2011. pp. 34-39

HSU C. W., CHANG C. C., LIN C. J.; **A practical guide to support vector classification.** Department of Computer Science, National Taiwan University, Taipei, Taiwan, 2003, 16 p.

JANSEN A., SCHULZ C. **Water demand and the urban poor: A study of the factors influencing water consumption among households in Cape town, South Africa.** South African Journal of Economics, Set. 2006, Vol. 74 (3), pp.593-609.

KOHAVI, R.; **A study of cross-validation and bootstrap for accuracy estimation and model selection.** International joint Conference on artificial intelligence. 1995. v. 14, pp. 1137–1145.

KULINKINA, A. V.; KOSINSKI, K. C.; LISS, A.; ADJEI, M. N.; AYAMGAH, G. A.; WEBB, P.; GUTE D. M.; PLUMMER, J. D.; NAUMOVA, E. N.; **Piped water consumption in Ghana: A case study of temporal and spatial patterns of clean water demand relative to alternative water sources in rural small towns.** Science of the Total Environment. Abr. 2016. Vol. 559. pp. 291–301.

LOUREIRO, D. **Metodologias de Análise de Consumos para a Gestão Eficiente de Sistemas de Distribuição de Água.** 2010. Dissertação (Doutorado em Engenharia Civil) – Instituto Superior Técnico, Universidade Técnica de Lisboa, Lisboa. Set. 2010. Disponível em: <repositorio.lnec.pt:8080/handle/123456789/1001450>. Acesso em: 09 de dez, 2016.

MARTINS, M. P. S. **Inovação Tecnológica e Eficiência Energética.** 1999. 43 F. Dissertação (Pós-Graduação em Energia Elétrica) – Instituto de Economia, Universidade Federal do Rio de Janeiro, Rio De Janeiro, 1999.

MEHTA, N.; DANG, S.; **Temporal Sequential Pattern In Mineração de Dados Tasks.** International Journal on Computer Science and Engineering. Jul. 2013. Vol. 3:7. pp. 2674-2678.

MOREIRA, T. O. M.; **Gestão de Perdas em Redes de Abastecimento de Água - assistida pelo simulador EPANET**. Dissertação – Faculdade de Engenharia, Universidade do Porto, Porto. 2011. Disponível em: < <https://repositorio-aberto.up.pt/bitstream/10216/65773/1/000146678.pdf> >. Acesso em: 15 de out. 2016.

MUTIKANGA, H. E.; SHARMA, S.; VAIRAVAMOORTHY K.; **Water Loss Management in developing countries: Challenges and prospects**. Journal AWWA. Dez. 2009. Vol. 101:12. pp. 57-68

NAGI, J., YAP, K. S., TIONG S. K., AHMED, S. K., MOHAMAD, M.; **Nontechnical Loss Detection for Metered Customers in Power Utility Using Support Vector Machines**. IEEE TRANSACTIONS ON POWER DELIVERY, abr. 2010, Vol. 25:2. PP. 1162-1171.

OLIVEIRA, F. M. R.; **REDUÇÃO DE PERDAS REAIS EM SISTEMAS DE ABASTECIMENTO DE ÁGUA: Otimização das pressões numa grande rede de distribuição de água – aplicação ao caso do Porto**. Dissertação (Mestrado em Hidráulica) – Faculdade de Engenharia, Universidade do Porto, Porto. 2013. Disponível em: < https://sigarra.up.pt/feup/pt/pub_geral.show_file?pi_gdoc_id=360844 >. Acesso em: 07 de dez. 2016.

PROCEL. PROGRAMA NACIONAL DE CONSERVAÇÃO DE ENERGIA ELÉTRICA – PROCEL. SANEAMENTO. **Procel Sanear**. 2015. Disponível Em: < <http://www.procelinfo.com.br/data/pages/lumis623fe2a5itemid6d82cf76dd284e7b8a607f31cb419a79ptbrie.htm> > acessado em 18 maio. 2015.

RATHOD, R. R.; GARG, R. D.; **Regional electricity consumption analysis for consumers using Mineração de Dados techniques and consumer meter reading data**. International Journal of Electrical Power & Energy Systems, Jun. 2016. pp.268-374.

REITERMANOVA Z.; **Data Splitting**. WDS'10 Proceedings of contributed papers, part I– Mathematics and Computer Sciences, Prague. 2010. pp. 31-36

SECRETARIA NACIONAL DE SANEAMENTO AMBIENTAL. **Abastecimento de Água: gerenciamento de perdas de água e energia elétrica em sistemas de abastecimento: guia do profissional em treinamento:nível 2**. Salvador, 2008. 139 p.

SHMUELI, G.; BRUCE, P. C.; STEPHENS, M. L.; PATEL, N. R.; **Mineração de Dados for Business Analytics: Concepts, Techniques, and Applications with JMP Pro**. Hoboken, USA; Abr. 2016. Disponível em <
https://books.google.com.br/books/about/Data_Mining_for_Business_Analytics.html?id=Ia87CwAAQBAJ&redir_esc=y>. Acesso Em: 08 dezembro. 2016.

SILVA, A.; ANTUNES, C.; **Multi-relational pattern mining over data streams**. Mineração de Dados and Knowledge Discovery. Nov. 2015. Vol. 29:6. PP.1783 -1814

SILVA, B.O., MONTEIRO, C.O., TORRES, C.V., SHINZATO, E., MOKARZEL, F.C., GUIBOSHI, M., PELLEGRINI, T.C. (2003). **Controle de Perdas de água em Sistemas de Distribuição**. Água em Ambientes Urbanos, 12 de Outubro, São Paulo, Brasil.

SOUZA, E. V.; COVAS, D. I. C. ; SOARES, A. K. . **Integrated Methodology for Enhancing Water and Energy Efficiency in Water Supply Infrastructures**. In: Water Loss 2010, 2010, São Paulo. Proceedings of the Water Loss 2010 Conference, 2010.

SPIRICA, J. F., DOČIĆB, M. B.; STANKOVIĆB, S. S.; **Fraud detection in registered electricity time series**. International Journal of Electrical Power & Energy Systems. Volume 71, out. 2015. pp. 42–50

TAVARES, G.R.; ZANOTELI, M. M.; D'ANGELLIS, T. da S. O.; SENA, L. da C. R.; **Efeitos da pesquisa de vazamentos na redução das perdas de água no município de Ponte Nova – MG**. In: XIX Exposição de Experiências Municipais em Saneamento, 2015, Poços de Caladas. Anais da XIX Exposição de Experiências Municipais em Saneamento, 2015, Poços de Caladas: ASSEMAE, 2015. P. 12. Disponível Em: <
<http://www.trabalhosassemae.com.br/sistema/repositorio/2015/1/trabalhos/64/405/t405t9e1a2015.pdf>>. Acesso Em: 22 de out. 2016.

TRATA BRASIL, **Saneamento Básico: Instrumento Urbano de Sustentabilidade**. 2014. Disponível em: <http://www.tratabrasil.org.br/saneamento-basico-instrumento-urbano-de-sustentabilidade>. Acesso em: 20 de out. 2015.

TSUTIYA, M. T. **Redução do Custo de Energia Elétrica em Estações Elevatórias de Água e Esgoto**. In: XIX Congresso Brasileiro de Engenharia Sanitária e Ambiental, 1997, Foz do Iguaçu. Anais do XIX Congresso Brasileiro de Engenharia Sanitária E Ambiental. Rio De Janeiro - RJ: Abes, 1997. P. 268- 269. Disponível Em: <<Http://Www.Bvsde.Paho.Org/Bvsacd/Abes97/Costo.Pdf>>. Acesso Em: 27 de mai. 2015.

VICENTINI, Liliana Pedroso. **Componentes do balanço hídrico para avaliação de perdas em sistemas de abastecimento de água**. 2012. Dissertação (Mestrado em Engenharia Hidráulica) - Escola Politécnica, Universidade de São Paulo, São Paulo, 2012. Disponível em: <<http://www.teses.usp.br/teses/disponiveis/3/3147/tde-03072013-151444/>>. Acesso em: 27 de out. 2015.

WANG, J.; CARDELL-OLIVER, R.; LIU, W.; **An incremental algorithm for discovering routine behaviours from smart meter data**. Knowledge-Based Systems. Dez. 2016, Vol. 113. pp. 61–74

XIN, K.; TAO, T.; LU, Y.; XIONG, X.; LI, F.; **Apparent Losses Analysis in District Metered Areas of Water Distribution Systems**. Water Resource Manage. Jan. 2014, Vol. 28. pp. 683-696.

YOUNG, R. **Saving Water and Energy Together: Helping Utilities Build Better Programs**. Alliance for Water Efficiency, out. 2013, 32 p.

YOUNG, R.; MACKRES, E.; **Tackling the Nexus: Exemplary Programs that Save Both Energy and Water**. Report E131 – ACEEE – American Council for Energy-Efficient Economy. Washington, DC. Jan. 2013. p. 111. Disponível em: <<http://aceee.org/node/3078?id=5129>> . Acesso em: 08 dez. 2016.

