ISSN 2359-4799
Volume 7 / Número 1 / Ano 2021 – p. 01-18
DOI: 10.36524/ric.v7i1.769

REVISTA ifes ciência

APLICAÇÃO DE TÉCNICAS DE QUALIDADE DA INFORMAÇÃO EM SENSORES NA INTERNET DAS COISAS (IoT)

APPLICATION OF INFORMATION QUALITY TECHNIQUES IN SENSORS ON THE INTERNET OF THINGS (IoT)

[1]Eduardo Esmínio Usbert
[2*]Allan Francisco Forzza Amaral
[3]Victório Albani de Carvalho

[1]Instituto Federal do Espírito Santo. E-mail: eusbert@msn.com
[2]Instituto Federal do Espírito Santo. E-mail: allanf@ifes.edu.br
[3]Instituto Federal do Espírito Santo. E-mail: victorio@ifes.edu.br
*Autor de correspondência

Artigo submetido em 26/07/2020, aceito em 24/03/2021 e publicado em 03/05/2021.

Resumo: Nas infraestruturas da Internet das Coisas é possível encontrar uma grande variedade de sensores de baixo custo, formando uma rede que gera um grande volume de dados, muitos deles anômalos e que podem impactar nas aplicações que os utilizam. Neste artigo demonstramos que, com algoritmos estatísticos e filtragem dos dados na borda da rede dos sensores, podemos identificar e eliminar esses dados anômalos. Através de um protótipo com dois sensores de temperatura, testamos dois algoritmos estatísticos de detecção de dados anômalos, que devidamente filtrados na borda, contribuem para o aumento da confiabilidade dos dados a serem armazenados e/ou consumidos por aplicações e/ou usuários diretamente. A Qualidade da Informação tem papel fundamental para esta confiabilidade e se relaciona com as diversas dimensões da avaliação, algumas das quais exploradas neste trabalho. Os resultados demonstram que esta avaliação detecta muitos dados coletados que podem ser descartados, contribuindo com as infraestruturas e aplicações que fazem o uso deles.
Palavras-chave: Internet das Coisas (IoT), Redes de Sensores Sem Fio (RSSF), *Outliers*, Qualidade da Informação (QoI).

Abstract: In the infrastructures of the Internet of Things, it is possible to find a wide variety of low-cost sensors, forming a network that generates a large volume of data, many of them anomalous and that can impact the applications that use them. In this paper we demonstrate that, with statistical algorithms and data filtering at the edge of the sensor network, we can identify and eliminate this anomalous data. Through a prototype with two temperature sensors, we tested two statistical algorithms for detecting anomalous data, which properly filtered at the edge, contribute to increase the reliability of the data to be stored and / or consumed by applications and / or users directly. Information Quality plays a fundamental role in this trust and is related to the different dimensions of evaluation, some of which are explored in this work. The results show that this evaluation detects a lot of collected data that can be discarded, contributing to the infrastructures and applications that make use of them.
Keywords: Internet of Things (IoT), Wireless Sensor Network (WSN), Outliers, Quality of Information (QoI)

1 INTRODUÇÃO

O progresso no campo de dispositivos embarcados permite que objetos do mundo real, por exemplo, eletrodomésticos, máquinas industriais, redes de sensores, celulares entre outros, se conectem à Internet. A consequência deste progresso pode ser notada com a criação de novos paradigmas, como a Internet das Coisas (IoT) (GUBBI *et al*., 2013). Sensores acoplados nestes objetos, isto é, embarcados em entidades físicas e conectados à Internet, criam novas oportunidades de projetos para aplicações interativas as quais conterão informação em tempo real referentes a lugares e objetos do mundo real (DE FRANÇA *et al*., 2011).

Esse é um contexto no qual as Redes de Sensores Sem Fio (RSSF) se destacam como uma importante forma de monitorar os fenômenos do mundo real. Essas redes são formadas por nós com, ainda que limitada, capacidade de processamento, armazenamento, comunicação, permitindo aos usuários observarem com nível de detalhes os ambientes ou entidades de interesse (por exemplo, nível de CO_2 de uma sala ou a umidade do solo numa área irrigada) e incorporam muitos dos novos conceitos de IoT.

Muitas pesquisas na área de IoT/RSSF culminam em diferentes tipos de aplicações, protocolos, arquiteturas e frameworks em que são apresentadas diferentes soluções para os problemas de restrições. Num primeiro momento, tais pesquisas concentraram-se no desenvolvimento de infraestruturas de suporte, incluindo algoritmos, protocolos, sistemas operacionais e linguagens, além de plataformas de hardware de sensoriamento. Protocolos MAC (SHAIK; ESWARAN, 2012), protocolos de roteamento (KO *et al*., 2011); algoritmos de disseminação, difusão e sincronização; projetos de linguagens e sistemas operacionais como nesC, TinyOS, Contiki (AKYILDIZ; VURAN, 2010); infraestruturas de suporte a IP, IPv6, 6LoWPAN (HUI; CULLER; CHAKRABARTI, 2009); arquiteturas de middleware (BHUYAN *et al*., 2014), e outras questões de infraestrutura direcionaram boa parte das pesquisas.

Mais recentemente temos observado um movimento em direção ao desenvolvimento de aplicações reais, geralmente tomando por base as infraestruturas de suporte desenvolvidas ao longo dos últimos anos, por exemplo, em projetos específicos tais como OpenIoT (KEFALAKIS *et al*., 2013), Almanac (BONINO *et al*., 2015) e MakeSense (CASATI *et al*., 2012). Cenários tais como *Smart Cities*, *Smart Home*, *Healthcare*, Logística, dentre outros cenários reais, vem sendo explorados nestes e outros trabalhos reportados na literatura.

Como essas arquiteturas tratarão os novos requisitos que surgem, particularmente no universo de IoT, são ainda um campo de pesquisa em andamento. Além de questões tradicionais, como a preocupação com a conservação energética e dos demais recursos dos nós das redes, outras questões vêm sendo bastante discutidas atualmente, tais como a qualidade dos dados coletados pelos sensores (BISDIKIAN; KAPLAN; SRIVASTAVA, 2013) e a introdução de mecanismos que facilitem a integração de IoT ao mundo dos processos de negócio e tomada de decisão (MEYER; RUPPEN; MAGERKURTH, 2013). Todas essas questões fundamentais precisam ser tratadas de maneira integrada pelas infraestruturas de suporte a IoT.

As infraestruturas de suporte, por sua vez, normalmente são construídas para tratar de problemas em diversos cenários nos quais o uso de IoT/RSSF se mostra como uma das possíveis alternativas para

controlar e monitorar ambientes. Dentre estes cenários é possível destacar desde aqueles mais amplos (por exemplo, *Smart Cities*) até os mais limitados (por exemplo, o controle da temperatura de uma sala). Independente do caso, questões sobre como avaliar a qualidade dos dados coletados pelos sensores é uma questão a ser considerada, uma vez que tais dados poderão ser armazenados e usados como fonte de pesquisas para tomada de decisão (por máquinas, sistemas ou pelo homem) ou para fins de registros históricos.

A questão é que, em muitos cenários, sensores podem produzir dados que podem não refletir a realidade (por exemplo, dado não preciso, não exato, desatualizado, incompleto) e, dependendo da sua tecnologia, também podem produzir dados anômalos.

Muitos destes cenários e problemas de aplicação podem ser melhorados através do tratamento das questões relacionadas à Qualidade da Informação (QoI), contribuindo para a melhoria de soluções propostas. Estas melhorias estão ancoradas em alguns elementos que podem ser integrados e utilizados para propor soluções em algumas áreas nas quais os sensores são utilizados.

A captura de fenômenos (por exemplo, temperatura, umidade, pressão) em entidades (pessoa, lugar ou objeto) requer que algumas preocupações sejam levadas em conta. Estas preocupações perpassam, por exemplo, desde o momento da captura do fenômeno até a disponibilização da informação para o consumidor dos dados. Consideramos este último argumento, dentre outros, um dos pilares deste trabalho: ao disponibilizar o dado para consumo de usuários ou aplicações precisamos, minimamente, passar um primeiro filtro e aumentar o grau de certeza de que algum processo de tomada de decisão está sendo feito baseado num dado previamente avaliado. Num segundo momento, podemos nos beneficiar desta abordagem de outras formas: a avaliação prévia do dado permite ganhos consideráveis para as infraestruturas, com destaque para a diminuição das bases de dados (e consequentemente menor atraso na recuperação de dados por aplicações) e a economia de recursos energéticos dos nós que compõem a rede de sensoriamento.

Sendo assim, o uso de técnicas e abordagens baseadas, por exemplo, em métodos estatísticos é bem útil para as infraestruturas e aplicações IoT e permite avaliar a qualidade dos dados coletados pelos sensores, promovendo confiabilidade em ambientes caracterizados pela alta heterogeneidade de hardware (sensores, dispositivos de comunicação e processamento).

Para tanto, os objetivos específicos serviram de guia para atingir as metas desta proposta:

- Implementar um protótipo para coletar, armazenar e avaliar dados coletados por sensores de temperaturas;
- Implementar e testar algoritmos de detecção de valores anômalos;
- Entregar dados avaliados para consumidores, sejam pessoas, máquinas ou aplicações;

1.1 DELIMITAÇÕES DA PESQUISA

Ao considerar os novos paradigmas da IoT ficamos expostos aos mais diversos cenários de aplicações. Muitos fenômenos físicos do mundo real podem ser capturados por meio de sensores e muitos deles ocorrem de forma espontânea, aleatória e caótica, tornando os cenários muito dinâmicos e com eventos inesperados. Essa abordagem procura se limitar a cenários mais controlados ou minimamente focados

naqueles em que se esperam menor dinamicidade, embora em muitos casos essa proposta possa ser testada e utilizada.

Uma vez que as técnicas de detecção de dados anômalos utilizadas neste trabalho não tem conhecimento prévio da condição de ocorrência, isto é, sua origem (ABID; KACHOURI; MAHFOUDHI, 2014), optou-se por identificar tal ocorrência por meio de comparações entre dados univariados coletados dos próprios sensores, justificando a abordagem não paramétrica e não supervisionada (YANG ZHANG; MERATNIA; HAVINGA, 2010). Além disso, essa abordagem não é adequada para cenários como aqueles de aplicações IoT em que o tempo de resposta é essencialmente crítico (*time-real*), uma vez que as leituras dos valores obedecem a um intervalo de tempo e uma condição pré-estabelecida.

Outra questão especialmente importante é sobre a origem dos dados. Esse trabalho não está focado especificamente na avaliação dos sensores, seus tipos, marcas, modelos ou fabricantes, tampouco nos algoritmos utilizados para avaliar os dados, embora seja inevitável tecer ao longo do trabalho comentários sobre suas características e comportamentos. A questão aqui é que sensores e algoritmos são usados como um meio para um fim, isto é, os dados produzidos por estes sensores que são objetos de estudo desta proposta.

Nas seções seguintes apresenta-se uma revisão de trabalhos da literatura sobre as teorias nas quais se embasou para abordar o tema e delimitar o escopo deste trabalho. Além disso, apresenta-se os métodos utilizados para avaliar os dados lidos pelos sensores e, em seguida, discute-se os resultados obtidos com os algoritmos de avaliação. Por fim, chegou-se as conclusões e os desdobramentos deste trabalho.

2 REFERENCIAL TEÓRICO

A Internet das Coisas (IoT – Internet of Things) apresenta um cenário no qual milhões de objetos são interconectados e dinamicamente integrados à Internet, compondo uma visão que alguns autores vem comparando a uma imensa pele digital (“digital skin”) (AL-FUQAHA *et al*., 2015). Os desafios para o desenvolvimento de infraestruturas de IoT englobam questões que vão além da tradicional adaptação dos protocolos e economia de recursos dos sensores. O controle da qualidade dos dados coletados passa a ser primordial, uma vez que decisões autônomas podem ser tomadas por máquinas, baseando-se em dados imprecisos dos sensores e/ou em informações equivocadas geradas nos sistemas de informação a partir dos dados sensoreados, que podem ter um impacto significativo nos processos de tomada de decisão e levar a consequências incalculáveis para pessoas, ambientes e negócios.

Isso é especialmente crítico quando bases de dados com valores errôneos podem ser compartilhadas por serviços e aplicações, em diferentes domínios, amplificando o problema. Neste sentido, o uso de teorias e estudos sobre *Quality of Information* (QoI) (SACHIDANANDA *et al*., 2010) assume grande relevância na pesquisa em IoT (BISDIKIAN *et al*., 2013). Por exemplo, a criação de abordagens baseadas em métodos estatísticos para avaliar as diversas dimensões, como a relevância e a confiabilidade dos dados coletados pelos sensores pode ser bem útil para as infraestruturas e, particularmente, para o universo de IoT, onde as soluções devem promover confiabilidade em ambientes caracterizados pela alta heterogeneidade de soluções de infraestrutura de hardware e software.

2.1 DIMENSÕES DE QoI

A qualidade da informação é naturalmente e empiricamente reconhecida

como importante para os consumidores de informação (LEE *et al*., 2002). Um ponto importante abordado por (BAQA *et al*., 2018) diz respeito à QoI como indicador de confiança dos usuários das aplicações. Sabe-se que apenas QoI não é suficiente para gerar a confiança de um usuário em uma aplicação ou serviço: outros fatores contribuirão, como as experiências anteriores e/ou a reputação do serviço. No entanto, como todos os aplicativos e serviços IoT dependem dos dados coletados, a QoI desempenha um papel fundamental na confiança entre usuários e serviços IoT. Nesse caso, constrói-se a confiança do dado relacionando-o diretamente com as diversas dimensões de avaliação da qualidade da informação. A seguir estão destacadas as descrições destes aspectos (multi) dimensionais compilados das definições em (FAGUNDES, 2017); (SACHIDANANDA *et al*., 2010); (BISDIKIAN; KAPLAN; SRIVASTAVA, 2013); (BAQA *et al*., 2018); (KAHN; STRONG; WANG, 2002):

- Acessibilidade: determina a medida na qual as informações estão disponíveis para serem acessadas;
- Suficiência: mostra se a informação é suficiente para as necessidades dos usuários;
- Interpretação Concisa: fusão entre representação concisa e interpretabilidade. Dispõe de uma análise clara, compacta e objetiva, isto é, sucinta e remida;
- Exatidão: valores próximos ao do mundo físico, podendo ou não ser/estar calibrados;
- Precisão: capacidade de reproduzir os mesmos valores dentro de uma faixa de desvio aceitável;
- Objetividade: é uma informação neutra, imparcial, direta;
- Credibilidade ou Confiabilidade: é a informação sem alteração da origem ao destino;
- Relevância: se as informações têm utilidade e aplicabilidade para determinada tarefa específica;
- Valor agregado: tem a ver com o benefício e vantagem oferecida pelo uso da informação;
- Perfeição: em que ponto de amplitude e profundidade as informações atendem a tarefa específica;
- Quantidade de dados: relevância da quantidade de informações adequada para cada tarefa;
- Facilidade de entendimento: se a informação é facilmente compreendida;
- Representação consistente: a que ponto as informações são apresentadas em mesmo formato;
- Atualidade: Trata da "idade" (tempo de vida) de um dado;
- Segurança/ Reputação/ Proveniência: cuida de onde vem o dado e preocupa-se com quem teve acesso a ele;
- Veracidade: dedica-se aos dados que podem ser convertidos em informações que atendam às qualidades necessitadas pelas aplicações utilizadas;
- Variabilidade: procede quanto a instabilidade do valor;
- Pontualidade: tempo para ser adquirido;
- Singularidade: dados livres de redundância;
- Reusabilidade: capacidade de ser usado novamente em situações distintas ou não.

Estas dimensões representam uma coleção de características da informação que permitem as aplicações decidirem se a informação é útil para seus propósitos. Estas características estão normalmente relacionadas aos dados no seu formato inteligível e/ou aos seus metadados, que complementam uma ou mais características daquele dado. Por exemplo, ao tratarmos as dimensões “Pontualidade” ou “Proveniência” de determinado dado estamos nos referindo as suas questões

temporais (quando foi coletado) e espaciais (onde foi coletado).

Por outro lado, ao tratarmos o teor do dado, estamos nos referindo ao valor compreensível de um fenômeno mensurável. Por exemplo, um sensor de temperatura que reporte o valor de 17º C ou um sensor reportando valores de 65% de umidade relativa, independente de contexto.Independentemente se a avaliação ocorrerá nos dados e/ou metadados, em ambos casos temos uma questão importante a ser considerada: o processo de avaliação das características da informação para provimento de QoI para as aplicações. Este processo determina o quão adequado é a informação para um uso particular, avaliando os dados contra um número de dimensões de qualidades já citados previamente (exatidão, disponibilidade, pontualidade, entre outros).

Se de um lado temos as diversas dimensões de QoI, que demonstram o grau de complexidade necessário ao avaliar um dado, do outro lado temos as técnicas utilizadas para avaliar as características destes dados lidos pelos sensores sob uma ótica multidimensional de QoI.

Considerando esta complexidade e também que os sensores na IoT normalmente têm recursos computacionais restritos, este trabalho limitou-se a avaliar o teor do dado por meio da análise de duas dimensões: Exatidão e Precisão. A questão aqui é investigar se valor dos dados de fenômenos reportados pelos sensores são confiáveis para serem armazenados e usados por aplicações. Para isso, aplicou-se técnicas da estatística descritiva para avaliar os dados sob as dimensões selecionadas, embora nem todas as dimensões possam ser avaliadas por meio destes métodos. A seção a seguir explora algumas destas técnicas e destaca aquelas usadas com maior frequência neste trabalho.

2.2 MÉTODOS ESTATÍSTICOS PARA BUSCA DE *OUTLIER*S

A IoT tem a missão de conectar o mundo real ao mundo virtual, tal como, por meio de sensores e atuadores. Sensores, por sua vez, convertem os sinais do mundo físico em dados passíveis de serem medidos e processados pelos sistemas computacionais. Porém, os sensores utilizados em boa parte de infraestruturas IoT são de baixo custo e, consequentemente, são suscetíveis a falhas. Portanto, uma grande quantidade dos dados coletados é imperfeita.

Para (LIMA, 2014), o custo e o tempo gasto em manutenções destas infraestruturas são, de certa forma, inviáveis devido à alta granularidade dos sensores. Além disso, dados defeituosos coletados pelos sensores podem chegar próximo a 50% em aplicações IoT (considerando falhas, além dos sensores, da infraestrutura de comunicação). É primordial melhorar a confiabilidade dos dados coletados e o reconhecimento imediato destas falhas colabora para o crescimento desta confiabilidade.

Alguns autores como (FAWZY; MOKHTAR; HEGAZY, 2013), (YANG ZHANG; MERATNIA; HAVINGA, 2010) e (ABID; KACHOURI; MAHFOUDHI, 2014) abordam a qualidade da informação com base no conjunto de dados produzidos por sensores IoT. Isto é feito através de técnicas estatísticas, *data mining*, inteligência artificial, aprendizado de máquina e teoria da informação. O termo comumente utilizado no âmbito dessas técnicas é *outlier*. Este termo tem suas origens na estatística e também é conhecido como anomalia. Um *outlier* é uma observação (ou um subconjunto de observações) que parece ser inconsistente com o resto do conjunto de dados. Da mesma forma, no universo IoT, *outliers* podem ser definidos como as medidas que se desviam significantemente do padrão

normal dos dados sensoreados (YANG ZHANG; MERATNIA; HAVINGA, 2010). Esta definição se baseia no fato de que os sensores têm a atribuição de monitorar o mundo físico que representa um padrão de comportamento normal nos dados sensoreados. Em situações práticas é comum que um ou possivelmente mais dados difiram do seu conjunto. Para tanto, técnicas estatísticas são utilizadas para decidir se os valores devem ou não ser rejeitados.

Em análises efetuadas em dados provenientes de sensores é possível identificar que, no processo de conversão de sinais, sistematicamente, podem ocorrer diversas falhas que podem produzir valores anômalos (NI *et al*., 2009) causados, por exemplo, por uma grande quantidade de ruído, erro de calibração, conexão ou hardware, bateria fraca e ambiente fora da faixa de medição do sensor.

Determinado sensor pode apresentar várias falhas simultâneas daquelas relacionadas previamente. Mesmo apresentando falhas, alguns dados gerados poderão ser informativos, enquanto os demais descartados. Cada ocorrência de falha pode gerar um impacto e, em alguns casos, os dados podem ser descartados totalmente, que é o caso dos *outlier*s, por não oferecerem nenhuma informação útil. Além disso, não eliminar os *outliers* traz consequências, uma vez que eles podem distorcer de forma significativa resultados estatísticos, como média, variância, gradiente, entre outros (LIMA, 2014).

A maioria dos trabalhos de destaque em detecção de *outlier*s foca na área da estatística, assunto que tem sido examinado por décadas. Dezenas de testes estatísticos podem ser utilizados para descobrir discordâncias/anomalias em dados usados nas aplicações. Estes testes se distinguem quanto ao tipo de distribuição dos dados, quanto ao tipo dos dados propriamente dito, quanto à natureza dos *outlier*s esperados, dentre outros (CAMPOS, 2015). Existem diversos modelos de distribuições teóricos na estatística que buscam traçar comportamento através de determinado evento em conformidade com sua ocorrência (CARMELO, 2018). Por tratar-se de tema muito importante e de interesse, o estudo de *outlier*s conquistou e continua a conquistar diversos pesquisadores em várias áreas de estudo. A descoberta de *outlier*s em amostras univariadas é um dos tópicos de grande importância na literatura estatística (GOLEMAN, DANIEL; BOYATZIS, RICHARD; MCKEE, 2019).

Dentre os testes mais comuns para detecção de *outliers* destacam-se o teste de Dixon, Chauvenet e *Grubbs* (GRUBBS, 1969) (OLIVEIRA, 2008); (LUCATO; COUTO; LUZ, 2007), além do teste dos Quartis (NISHA *et al*., 2014). Dentre estes, selecionamos o teste dos Quartis e o teste de Grubbs para utilizarmos neste trabalho, uma vez que melhor se adequam aos objetivos do projeto pela baixa complexidade e custo computacional.

Aritmeticamente simples, o teste dos Quartis divide um conjunto de dados ordenados (em ordem crescente) em três partes iguais (Q1, Q2 e Q3). O primeiro quartil (Q1) divide os 25% dos valores mais baixos dos 75% que são maiores do que eles na amostra. O segundo quartil (Q2) representa a mediana, ou seja, 50% dos valores são menores que a mediana e 50% dos valores são maiores. O terceiro quartil (Q3) divide a parcela correspondente dos 75% dos valores mais baixos dos 25% dos valores que são maiores do que eles. Dessa forma, a Amplitude Interquartil (Ai) pode ser definida como Ai = a3(Q3) – a1(Q1), onde a3 representa o valor da amostra referente ao terceiro quartil (Q3) e a1 representa o valor da amostra referente ao primeiro quartil (Q1). *Outlier*s podem ser detectados numa amostra utilizando Amplitudes Superiores (Asup) ou Amplitudes Inferiores (Ainf), que são determinadas utilizando uma barreira *b*.

A barreira (*b*) indica o grau de dispersão que se deseja aplicar ao conjunto de dados para se detectar um valor considerado como *outlier*. Conforme a literatura da estatística descritiva, a barreira pode ser um número real, assumindo qualquer valor, mas que normalmente são utilizados dois valores: (*b* = 1,5) ou (*b* = 3). Ao utilizar a barreira 1,5 é possível captar mais de 99% dos dados abaixo/acima de uma curva normal (chamado de *outlier*s moderados), considerado, neste caso, como barreiras internas. A outra opção multiplica o valor lido (amostra) por 3. Neste caso, considera-se como barreiras externas e possibilita detectar *outlier*s extremos (valores muito acima do valor tido como normal para a série testada).

Dessa forma, a Amplitude Inferior é determinada por:

$$A_{inf} = a_1 - (b \times A_i)$$

Já a Amplitude Superior é dada por:

$$A_{sup} = a_3 + (b \times A_i)$$

Assim, uma amostra testada (a_t) é considerada um *outlier* quando $a_t < A_{inf}$ ou $a_t > A_{sup}$.

Nos nossos cálculos deste trabalho considerou-se *outlier*s extremos (*b* = 3), pois com os dados coletados e os testes efetuados com a amplitude apresentada, obteve-se melhores resultados com esta abordagem.

O teste de Grubbs (GRUBBS, 1969) é outro teste de detecção de *outlier*s recomendado pelo *International Organization for Standardization* (ISO). Este teste compara a distância, medida em desvios-padrão, do valor suspeito em relação à média do conjunto de valores (o valor suspeito é incluído no cálculo da média e do desvio-padrão). Se a distância for maior do que o limite crítico tabelado, o valor suspeito é considerado um *outlier*. Os valores críticos se referem à um índice na tabela de Grubbs (GRUBBS, 1969), que leva em consideração o grau de significância da amostra testada com relação ao total de amostras. O teste de Grubbs é dado por:

$$G = \frac{|a_t - \bar{a}|}{s}$$

A variável a_t representa a amostra testada, $\bar{a}$ representa a média e *s* o desvio padrão. Por sua vez, o desvio padrão é obtido por:

$$\sqrt{\left(\frac{\sum_{i=1}^{p}(ai - \bar{a})^2}{p}\right)}$$

Neste caso, *p* representa o número de elementos da amostra. Dessa forma, o valor G é calculado e comparado ao índice de significância α presente na tabela do índice de Grubbs. Sendo G o valor de Grubbs obtido e G_c o índice crítico, a amostra testada a_t é um *outlier* se $G > G_c$.

As técnicas de detecção de *outliers* com base em estatísticas são consideradas melhores para adaptação em redes de sensoriamento de baixa capacidade devido à sua escalabilidade, pelo baixo custo computacional e pela opção de não precisar de conhecimento dos dados *a priori*. Além disso, é possível destacar duas formas de detecção de *outlier*s em aplicações IoT: online, que é um processo executado imediatamente após a leitura e offline, efetuada após os dados serem enviados para uma base de dados (neste caso, a detecção é efetuada um tempo após o registro). A detecção online, além de manter a integridade dos dados, diminui o tempo para detecção, o que torna esta forma mais atrativa para aplicações que demandam leituras cujo tempo de resposta é uma questão a ser considerada (BINTI IDA UMAYA, 2017).

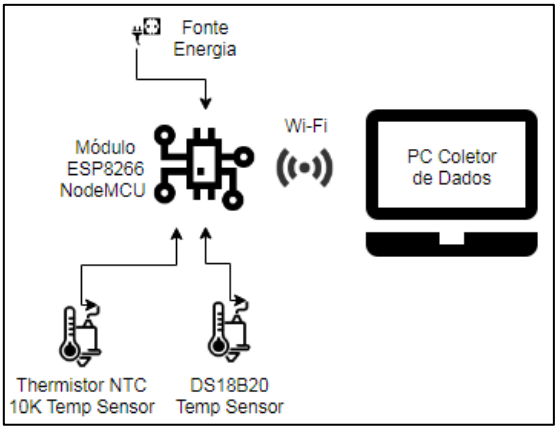
Um dos grandes desafios enfrentado pelas técnicas de detecção de *outlier*s em redes de sensoriamento nas infraestruturas de IoT é satisfazer os requisitos na mineração dos dados mantendo o consumo de recursos dos nós da rede ao mínimo. A questão principal é como processar tantos dados quanto possível de forma descentralizada mantendo baixa sobrecarga de comunicação, memória e processamento (AMARAL, 2016). Ainda que uma grande quantidade de dados seja gerada por medição dos sensores, o maior consumo de energia não está neste processo e sim na transmissão contínua desses dados. Portanto, uma forma de diminuir o consumo de energia é atenuar as transmissões de dados. Várias propostas de tratamento de dados com o intuito de minimizar o consumo de energia das redes de sensoriamento foram sugeridas na literatura. Ademais, a abordagem de diminuir o número de transmissões de dados ganhou grande atenção por aumentar a vida útil da rede e dos nós que a compõem (FATHY *et al*., 2018).

Outra consequência desta abordagem é a redução de bases de dados armazenadas e compartilhadas, especialmente nas infraestruturas IoT, que utilizam a nuvem como forma de obter dados. Isto implica em dizer que ao aplicar os algoritmos, as bases de dados serão resumidas de forma significativa e, ao mesmo tempo, mantendo dados úteis para as aplicações. Esta é uma das técnicas muito aplicada em tecnologias de Big Data (OUSSOUS *et al*., 2018) e tem por objetivo reduzir dados e a complexidade computacional utilizada para recuperar informações.

3 MATERIAIS E MÉTODOS

Esta seção explica os processos utilizados para coletar os dados dos sensores e avalia-los sob a ótica de duas dimensões: exatidão e precisão. O passo inicial consiste na montagem da infraestrutura de sensoriamento composto por um equipamento embarcado ESP8266 NodeMCU, no qual estão acoplados dois sensores de temperatura (um digital – modelo DS18B20; e outro analógico – modelo Thermistor NTC 10k), conforme diagrama da Figura 1.

Figura 1–Diagrama do protótipo utilizado para coleta de dados



O sensor analógico Thermistor NTC 10K encapsulado é muito utilizado em projetos de microcontroladores, devido ao seu baixo custo, porém, necessita ser calibrado dada a baixa exatidão. Além disso, sua precisão também não é satisfatória. Já o sensor digital DS18B20 (também encapsulado) possui, dentre outras vantagens, maior precisão e exatidão se comparado ao sensor analógico testado. O fato de usarmos dois tipos de sensores de temperaturas diferentes é para compararmos as margens de erros de leituras entre ambos os sensores, além de observarmos também as questões referentes as anomalias e leituras incorretas. Todo este processo de avaliação foi divido em três etapas, descritas a seguir.

Na primeira etapa, um setup foi construído para monitoramento de temperatura do ambiente, alternando entre ambiente climatizado a 22° C (em horário comercial) e não climatizado (temperatura ambiente -- horário não comercial), gerando dados reais num ambiente conhecido para melhor análise e estudo.

Por meio deste protótipo coletou-se dados de temperatura do sensor digital DS18B20 no período de 11/04/2019 a 03/07/2019, totalizando 140.730 registros. No período de 16/04/2019 a 03/07/2019 coletou-se dados de temperatura do sensor analógico Thermistor NTC 10k, totalizando 128.139 registros armazenados deste sensor. Nesta parte do trabalho testou-se empiricamente os dados coletados pelos sensores. Com a observação, análise e acompanhamento destes registros, identificou-se a importância de filtrar estas informações no próprio equipamento embarcado mesmo antes de serem enviados para o registro na base de dados. Isso possibilitou determinar a quantidade de amostras testadas e o tempo médio entre as leituras, questões consideradas importantes para os propósitos deste trabalho.

Na segunda etapa, desenvolveu-se e embarcou-se no protótipo dos sensores dois algoritmos de busca de valores extremos amplamente conhecidos na literatura estatística: o teste dos Quartis e o teste de Grubbs. Estes algoritmos foram utilizados para filtrar os dados a serem enviados para a base de dados.

Iniciou-se os testes com algoritmos de cálculos estatísticos registrando todas as medições, porém, marcando o *status* específico de cada uma. Ao enviar um dado a ser registrado na base de dados, o algoritmo faz uma avaliação prévia do valor e o marca com um dos 4 *status* criados, que podem ser: *a) Normal* - valores a serem transmitidos, pois estão dentro da curva da "normalidade" dos valores coletados; *b) Outlier* - valores "anômalos" que se apresentaram fora dessa curva, detectados pelo teste estatístico; *c) Repetida* - medição repetida que não apresentou nenhuma variação; *d) Keep-alive* - status criado para ser transmitido após longas séries de repetições e *outliers*, informando que o dispositivo embarcado continua em funcionamento. Nestes testes foram reunidos mais de 37 mil registros de cada sensor, de 30/09/2019 a 16/10/2019 e serviu de base de testes para estabelecer parâmetros, por exemplo, de amostras a serem testadas, a frequência de leitura e o comportamento da saída dos algoritmos.

3.1 REGISTRO DE DADOS AVALIADOS

Na terceira e última etapa de testes e avaliação foram coletados cerca de 15 mil registros de cada sensor com o teste estatístico dos Quartis e cerca de 15 mil registros com o teste estatístico Grubbs, totalizando aproximadamente 60 horas de testes. No teste estatístico dos Quartis considerou-se a coleta de 21 amostras antes de executar o algoritmo de avaliação. A mesma abordagem foi utilizada no teste estatístico de Grubbs. Para ambas as coletas, foi utilizado um intervalo médio de frequência de 15 segundos entre uma medição e outra.

Nossa base de dados está registrada em duas guias da planilha do *Google Sheets*, uma para cada sensor. A programação embarcada no módulo envia as informações automaticamente para um formulário do Google que, por sua vez, preenche a planilha vinculada, conforme a Figura 2. A vantagem do preenchimento da planilha através do formulário é que o primeiro campo "Carimbo de data/hora", que é pré-definido pela própria programação do Google, registra a data e a hora do evento, momento em que os dados são gravados na planilha.

Para os propósitos deste estudo, todas as leituras foram registradas no banco

Figura 2 - Imagem da planilha da base de dados

	A	B	C	D	E	F
1	Carimbo de data/hora	Temperatura	Hora de Envio	Status	Data Registro	Hora Registro
1589	25/10/2019 02:15:37	26.79	02:15:36	Repetida	25/10/2019	02:15:37:16
1590	25/10/2019 02:15:55	26.79	02:15:53	Repetida	25/10/2019	02:15:54:65
1591	25/10/2019 02:16:12	26.79	02:16:10	Repetida	25/10/2019	02:16:11:53
1592	25/10/2019 02:16:29	26.79	02:16:28	Keep-Alive	25/10/2019	02:16:28:97
1593	25/10/2019 02:16:47	26.79	02:16:45	Repetida	25/10/2019	02:16:46:56
1594	25/10/2019 02:17:04	26.70	02:17:03	Outlier	25/10/2019	02:17:03:70
1595	25/10/2019 02:17:21	26.79	02:17:20	Normal	25/10/2019	02:17:21:46

de dados com um dos *status* (Normal, O*utlier*, Repetida ou K*eep-Alive*), conforme explicado anteriormente e demonstrada na coluna *Status* da Figura 2. Observou-se que diversos registros da temperatura eram repetidos em sequência, sendo desnecessário o registro no banco de dados numa situação real, podendo ser descartada no próprio módulo sensor. Como trabalhou-se em ambiente controlado, pode-se observar que as repetições apresentavam a realidade do ambiente e que, além destas repetições, aconteciam variações ao longo do período, não sendo um erro de medida do sensor. Entretanto, para os propósitos da análise, efetuou-se esse registro com o status *Repetida*. As leituras que eram identificadas como *outlier*s e poderiam ser descartadas em uma situação real foram registradas com o status *Outlier*. As leituras que estavam em conformidade e que deveriam ser transmitidas e gravadas no banco de dados foram registradas com o status *Normal*.

Como numa situação real, na ausência de longo tempo sem registro de algum dado com o status *Normal*, optamos por sinalizar um *status* denominado *Keep-Alive*. Este *status* é enviado em caso de uma longa sequência do *status Outliers* ou *Repetida*. Isso é necessário em caso de dúvidas do funcionamento do módulo. Portanto, a cada 10 (dez) leituras – equivalente a aproximadamente 2,5 minutos – registradas como *Repetidas* e/ou *Outlier*s – é efetuada uma transmissão (independente de avaliação) cujo dado é registrado com o status *Keep-Alive*, indicando que o módulo continua coletando dados e transmitindo na rede.

3.2 PARÂMETROS UTILIZADOS NOS ALGORITMOS EMBARCADOS

O primeiro teste para identificar *outlier*s foi o teste estatístico dos Quartis. Neste algoritmo criou-se um vetor de 21 posições ao inicializar o módulo embarcado. Esta quantidade de posições no vetor foi determinada com base em testes empíricos com quantidades diferentes (descrito na primeira etapa de coleta de dados), que possibilitou identificar o espaço amostral adequado para obter melhores resultados. Este vetor é preenchido pelas amostras lidas pelo sensor de temperatura com medições a cada 15 segundos. Este processo ocorre para ambos os sensores utilizados no protótipo. Depois de todas as posições preenchidas, o vetor é colocado em ordem e é efetuado o teste dos Quartis. Importante salientar que este teste divide a amostra em 2 partes: o Q1, que é o primeiro Quartil e o Q3, o terceiro Quartil. Um quartil é, na verdade, uma posição do vetor que armazena uma leitura (uma amostra da temperatura). Com base no valor contido na posição é realizado um cálculo aritmético simples que multiplica o valor da leitura pela barreira (b = 3). Assim, qualquer valor abaixo de Q1 ou acima de Q3 é considerado um *outlier*. Depois do vetor preenchido, uma nova medição (também com a frequência de 15 segundos entre elas) é armazenada em uma posição diferente do

vetor, sempre eliminando a medição mais antiga, e novamente o vetor é colocado em ordem antes dos novos cálculos para identificar *outlier*s. Todo esse processo ocorre a cada nova leitura, isto é, o teste dos quartis sempre é executado quando chega um novo valor no vetor. A seguir um trecho do algoritmo que demonstra o cálculo dos quartis e testa se o valor é um *outlier*.

```
for (byte i = 0; i < tVet1; i = i + 1)

{

 qI1 = checkOLorder1[round(((tVet1 + 1) / 4))-1];        //Calcula o 1º Quartil, já arredondando, caso necessário.

 qIII1 = checkOLorder1[round((3 * (tVet1 + 1) / 4))-1];   //Calcula o 3º Quartil, já arredondando, caso necessário.

}

ampl1 = qIII1-qI1;     //Calcula a Amplitude

Serial.println();

Serial.print("1º Qualtil: ");

Serial.println(qI1);

Serial.print("3º Qualtil: ");

Serial.println(qIII1);

Serial.print("Amplitude: ");

Serial.println(ampl1);

lim1 = ampl1 * bar; // Multiplicando a amplitude pela barreira, para encontrar os limites.

Serial.print("Tamanho Vetor: ");

Serial.println(tVet1);

if (ampl1 > 0.0f)

{

 if ((checkOL1[cont] < ((qI1 - lim1)) || checkOL1[cont] > ((qIII1 + lim1))) && contKaL1 < kalLim) //Testa para verificar se a medição está abaixo ou acima dos limites e seta o status outlier caso sim.

 {

  status1 = "Outlier";

 }
```

O outro teste de *outlier*s aplicado é o teste estatístico de Grubbs. Neste algoritmo também criou-se um vetor com 21 posições, conforme realizado no algoritmo do teste dos Quartis, procurando manter um padrão entre os dois testes (ainda que o cálculo aritmético entre ambos seja diferente). O preenchimento inicial é efetuado da mesma forma, assim como a frequência de medição. Neste teste não há a necessidade de colocar o vetor em ordem. O teste de Grubbs também é um cálculo aritmético simples. O valor testado é subtraído da média das amostras e depois dividido pelo desvio padrão amostral. O valor resultante é comparado em uma tabela que indica se o valor testado é um *outlier* ou não.

Segue abaixo a parte do algoritmo que calcula o teste de Grubbs e testa se o valor é um *outlier*. Vale lembrar que o *indice* (variável do algoritmo) é um valor constante da tabela do índice de Grubbs para a quantidade de amostra testada.

```
for (byte i = 0; i < tVet1; i = i + 1)

{

 Serial.print(" - ");

 Serial.print(checkOL1[i]); //Apresenta no Monitor Serial o vetor.

}

soma1 = 0.0f;

for (byte i = 0; i < tVet1; i = i + 1)

{ //SOMA OS REGISTROS DO VETOR

 soma1 += checkOL1[i];
```

```
    }

        Serial.println();

    Serial.print("Soma1: ");

    Serial.println(soma1);

    media1 = soma1 / tVet1; // CALCULA A MÉDIA

    Serial.print("Média1: ");

    Serial.println(media1);

    somaDesv1 = 0.0f;

    for (byte i = 0; i < tVet1; i = i + 1)

    { //SOMA PARA CALCULAR O DESVIO PADRÃO

      somaDesv1 += (pow((checkOL1[i] - media1),2));

    }

    Serial.print("Soma para Desvio: ");

    Serial.println(somaDesv1);

    desvP1 = (sqrt(somaDesv1 / tVet1)); //CALCULA O DESVIO PADRAO

    Serial.print("Desvio Padrão: ");

    Serial.println(desvP1);

    Serial.print("Tamanho Vetor: ");

    Serial.println(tVet1);

    if (desvP1 != 0.0f)

    { //SE DESVIO PADRÃO NÃO FOR ZERO, CALCULA O ÍNDICE DE GRUBBS

      grubbs1 = ((fabs(temp1-media1)) / desvP1); // CALCULA O ÍNDICE DE GRUBBS

      Serial.print("Índice de Grubbs: ");

      Serial.println(grubbs1);

      Serial.print("Índice Crítico 10%: ");

      Serial.println(indice);

      if (grubbs1 > indice  && contKaL1 < kalLim) //Testa para verificar se a medição está abaixo ou acima dos limites e seta o status outlier caso sim.

      {

        status1 = "Outlier";

      }
```

Para ambos os algoritmos, depois de identificado o *outlier* o valor é, no nosso caso, armazenado na base de dados com o devido *status*. Em uma situação real, este dado poderia ser descartado.

Importante destacar que a ocorrência de *outlier*s é aleatória, normalmente causada por uma falha no sensor, problema de qualidade do dispositivo de sensoriamento ou outra causa já mencionada previamente. Portanto, para verificar o comportamento dos algoritmos na presença de falhas, ambos sensores foram induzidos a gerarem valores bem acima ou abaixo do esperado para as leituras em sequência, provocando o surgimento de *outlier*s. Isso é feito aproximando os sensores de temperaturas em superfícies mais frias ou mais quentes do que o ambiente no qual ele está inserido, provocando leituras bem diferentes do esperado.

Outro destaque é que a escolha e o desenvolvimento destes dois algoritmos não têm por objetivo avalia-los, ou seja, identificar o que melhor se destaca, mas sim de compararmos e verificar se as margens de erros de saída de ambos têm discrepâncias consideráveis a ponto de não podermos utiliza-los para os propósitos de nossos testes.

4 RESULTADOS E DISCUSSÃO

A leitura de dados de fenômenos físicos e químicos do mundo real, como temperatura, umidade, pressão, velocidade, gases, radiação infravermelha, luminosidade, dentre outros, normalmente geram uma grande quantidade de dados, variando de acordo com o intervalo de medições requerido. Em princípio, todas as amostras coletadas pelos sensores devem ser consideradas (pois podem apontar uma variação do ambiente monitorado), mas nem todas são úteis para a aplicação por não apresentarem alterações significativas. Nesse caso, pode-se economizar recursos da rede, dos sensores e dos servidores de armazenamento evitando o encaminhamento desse dado desnecessário até a aplicação final. Por outro lado, ainda que o dado seja importante para a aplicação, é necessário considerar que ele pode ser resultado de um erro de medição, o que pode levar o sistema a desencadear alertas falsos. Em ambos os casos, verificar a qualidade da informação coletada torna-se um requisito da aplicação, o que pode ser feito avaliando-se atributos relacionados com algumas das dimensões dos dados coletados. Essa avaliação pode empregar técnicas estatísticas dentro da própria rede de sensores (permitindo economia na transmissão de dados), quanto na aplicação (permitindo economia de processamento dentro da rede e maior independência entre a aplicação e a rede de sensores).

Nesse estudo optou-se por embarcar os algoritmos dentro do sensor, isto é, a abordagem *in-network*. Neste caso, os algoritmos estão embarcados de forma independente, ou seja, na borda da rede. Isso foi possível graças a simplicidade do funcionamento dos testes estatísticos, que consomem pouco processamento e memória, itens que normalmente são bem limitados na maioria dos nós de sensoriamento.

Neste projeto, como foi efetuada a coleta em ambiente controlado, foi utilizado um termômetro onde há um ar condicionado funcionando em tempo integral na temperatura de 22º C. Isso permitiu observar (empiricamente) que os sensores estavam coletando dados bem próximos à realidade (fenômeno físico do mundo real), que em relação à QoI refere-se à avaliação da *exatidão* do dado. Em face desta observação, está implícito que a avaliação da exatidão já acontece naturalmente ao buscamos valores *outliers* com os algoritmos. Já em relação a *precisão*, as leituras sequenciais iguais (idênticas) demonstram o grau de precisão do sensor, mesmo desprezando suas margens de erro (normalmente ± 0,2).

Não necessariamente afirma-se ou sugere-se, aqui, que os testes estatísticos empregados nesta pesquisa sejam ideais e/ou absolutos. Nem mesmo que se apliquem e solucionem todas as situações que necessitem de filtragem de dados na borda (módulo de sensoriamento / RSSF). Na verdade, demonstra-se que a filtragem dos dados na borda pode ser útil e necessária para contribuir para a confiabilidade dos dados.

Para os propósitos dos testes, 30 intervenções foram efetuadas nos protótipos com o objetivo de provocar os *outlier*s. Observando o levantamento dos dados das tabelas apresentadas é possível verificar na coluna Status *Outlier* da Tabela 1 (que

Tabela 1: Aplicação do Algoritmo dos Quartis

Sumarização dos Testes com o Algoritmo dos Quartis							
Sensor de Temperatura	Registros Coletados	Status Normal	Status Repetido	Status *Outlier*	Status Keep-Alive	*Outlier*s Falsos Positivos	Registros Descartáveis (%)
DS18B20	16.068	3.546	11.973	40	509	10	74,76%
NTC 10K	15.978	7.245	8.477	137	119	107	53,91%

apresenta os dados obtidos com o algoritmo dos quartis, através do sensor DS18B20), a incidência de 40 *outlier*s. Desse total, 10 são falsos positivos. Com o algoritmo em funcionamento efetuou-se os testes de intervenções colocando o sensor em contato com superfícies com temperatura maior ou menor do que a do ambiente que vinha sendo medida e armazenada no vetor de testes. Imediatamente, a temperatura era lida, armazenada em um dos campos do vetor e o teste devidamente efetuado. Esta ação permitiu a detecção de 30 leituras *outliers*. Vale ressaltar que em todas as intervenções o algoritmo reagiu imediatamente, identificando a leitura como *outlier*. Ao retirar o contato do sensor com a superfície de teste, as medições começavam imediatamente voltar ao normal. Os registros falsos positivos apresentam temperatura registrada aparentemente dentro do padrão, mas que pelo funcionamento do algoritmo recebeu o status *outlier*.

NTC 10K), observou-se uma particularidade: o sensor analógico é pouco preciso e o retorno às medições anteriores aos testes não eram tão imediatas e, às vezes, por um bom tempo, se mantinha com as medições mais próximas das intervenções do que das medições em condições da temperatura ambiente controlada. A suposição é que o tempo de resposta para restabelecer as medições da temperatura ambiente do sensor NTC é maior, causando um número maior de falsos positivos, neste caso, 107 leituras. Entretanto, uma vez que questões intrínsecas do funcionamento do sensor não fazem parte deste estudo e sim os dados produzidos por ele, não comprovou-se esta hipótese.

Sobre a Tabela 2, agora com o Algoritmo de Grubbs -- tanto para o sensor DS18B20 quanto para o NTC 10K – realizou-se os mesmos testes de intervenções, tendo as respostas esperadas tanto do algoritmo quanto dos sensores, que

Tabela 2: Aplicação do Algoritmo de Grubbs

Sumarização dos Testes com o Algoritmo de Grubbs							
Sensor de Temperatura	Registros Coletados	Status Normal	Status Repetido	Status *Outlier*	Status Keep-Alive	*Outlier*s Falsos Positivos	Registros Descartáveis (%)
DS18B20	19.538	4.797	13.432	693	616	663	72,30%
NTC 10K	19.477	10.442	8.304	670	61	640	46,07%

É importante ressaltar que este teste provoca um status *outlier* pela mudança repentina da temperatura medida pelo sensor, mas caso o teste se mantivesse por tempo prolongado, à medida que preenchesse os elementos do vetor, essas medições passariam a ter registros com o status normal. Esse é um comportamento esperado do algoritmo: como os *outlier*s são leituras anômalas e isoladas de valores distantes dos padrões esperados, o que ocorre normalmente é uma leitura anômala e a leitura seguinte volta a um valor próximo dos padrões esperados. Também na Tabela 1, (que também apresenta os dados obtidos com o algoritmo dos Quartis -- agora em relação ao sensor Thermistor

imediatamente acusava *outlier*. Nas intervenções, todos os 30 registros *outliers* foram identificados pelos sensores DS18B20 e NTC, além dos falsos positivos, que neste caso apresentam uma quantidade considerável. Para esta situação específica, considerou-se o teste de *Grubbs* com elevado grau de sensibilidade, podendo ser aplicado em cenários e situações em que uma pequena variação na amostra seja potencialmente significativa.

Pelas informações sumarizadas nas duas tabelas, grande parte dos dados não seria transmitida em uma situação real. Esse fato é comprovado pela coluna “Registros Descartáveis” que apresenta o percentual de dados coletados que não seriam

transmitidos, que neste caso é uma somatória dos dados classificados como *outliers* com as medições classificadas como repetidas.

Em específico, as medições repetidas refletem a avaliação da dimensão *precisão* do sensor, que gera valores idênticos num intervalo de tempo muito pequeno, podendo, então, ser descartados, gerando uma economia de 74,76% no envio de registros. No pior cenário, 46,07% não seriam transmitidos, configurando praticamente metade dos dados.

Se por um lado temos falsos positivos, do outro temos algoritmos que identificam imediatamente todas as intervenções efetuadas ao classifica-las como *outliers*. Além disso, também identificou-se e filtrou-se as medições repetidas, que proporciona benefícios consideráveis como a economia de recursos energéticos dos equipamentos da RSSF e aumento da vida útil dos sensores.

Por fim, os resultados das Tabelas 1 e 2 demonstram principalmente que através de testes estatísticos simples é possível efetuar a filtragem dos dados na borda de sensores presentes na IoT, contribuindo para o aumento da confiabilidade dos dados coletados por sensores. Isso diminui o excesso de informações desnecessárias nos bancos de dados que, com maior confiabilidade, podem ser consumidos diretamente pelas aplicações e pelos usuários.

5 CONCLUSÕES

O principal objeto de estudo deste trabalho é o aumento da confiabilidade dos dados coletados por sensores. O reconhecimento de dados anômalos colabora para o crescimento desta confiabilidade, dando foco a identificar e descartar outliers e medições repetidas já na borda dos dispositivos de sensoreamento.
Observou-se que os algoritmos estatísticos utilizados aqui atendem à realidade deste projeto, apesar do alto grau de sensibilidade apresentado pelo teste de Grubbs, como explicado na seção anterior. Porém, estes mesmos algoritmos não necessariamente são os mesmos que atenderão outras situações, deixando margem para outras pesquisas, testes e estudos. Também em relação à QoI, o foco foi direcionado à confiabilidade. Para tanto, avaliou-se as dimensões da exatidão, precisão e temporalidade, embora esta última não tenha sido usada para os propósitos desta pesquisa.

Neste experimento trabalhou-se com um módulo e dois sensores com os quais teve-se a oportunidade de acompanhar o grande volume de informações transmitidas e armazenadas num intervalo de tempo considerável. Através de filtragens simples na borda, é possível deixar de transmitir e de armazenar dados anômalos, que são resultados de erros de medição. Se transportarmos esta realidade para o mundo real da IoT, por exemplo Cidades e Casas Inteligentes, respeitadas as proporções e falarmos das devidas filtragens nas bordas de um grande universo de sensores, estaremos falando de uma economia de energia e espaço de armazenagem de dados de grandes proporções.

O universo IoT, especialmente com o foco de filtragem na borda, ainda é um campo de pesquisa em andamento preocupado não só com questões tradicionais como protocolos, padrões dentre outros, mas com outras questões como a qualidade dos dados coletados pelos sensores. A confiança se relaciona diretamente com as diversas dimensões de avaliação da qualidade da informação, que tem aspectos multidimensionais a serem ainda explorados. Nosso argumento é que QoI desempenha um papel fundamental na confiança entre usuários e serviços IoT, abrindo inúmeras possibilidades de novos estudos, observando-se as diversas dimensões da QoI. Este trabalho pode ser continuado através da análise de outros testes estatísticos, buscando eliminar ao máximo os outliers e falsos positivos. O

setup dos testes também pode ser melhorado e ampliado adicionando mais equipamentos embarcados num ambiente real não controlado, permitindo avaliar o comportamento dos sensores, algoritmos e módulo embarcado.

REFERÊNCIAS

ABID, Aymen; KACHOURI, Abdennaceur; MAHFOUDHI, Adel. Anomaly Detection in WSN : critical study with new vision. **International Conference on Automation, Control, Engineering and Computer Science (ACECS'14) Proceedings**, p. 1–9, 2014.

AKYILDIZ, Ian F.; VURAN, Mehmet Can. **Wireless Sensor Networks**. Chichester, UK: John Wiley & Sons, Ltd, 2010.

AMARAL, Allan Francisco Forzza. **Uma Proposta de Arquitetura de Redes de Sensores sem Fio Aplicada ao Monitoramento Térmico da Rede de Frios.** 2016. 128 f. Universidade Federal do Espírito Santo, 2016.

BAQA, Hamza *et al*. Quality of Information as an indicator of Trust in the Internet of Things. ago. 2018, [S.l.]: IEEE, ago. 2018. p. 204–211

BHUYAN, Bhaskar *et al*. A Survey on Middleware for Wireless Sensor Networks. **Journal of Wireless Networking and Communications 2014**, v. 4, n. 1, p. 7–17, 2014.

BINTI IDA UMAYA. DETECÇÃO E IDENTIFICAÇÃO DE OUTLIERS EM REDES DE SENSORES SEM FIO DE LARGA ESCALA. **Universitas Nusantara PGRI Kediri**, v. 01, p. 1–7, 2017.

BISDIKIAN, Chatschik; KAPLAN, Lance M; SRIVASTAVA, Mani B. On the Quality and Value of Information in Sensor Networks. **ACM Transactions on Sensor Networks (TOSN)**, v. 9, n. 4, p. 1–26, 2013.

BONINO, Dario *et al*. ALMANAC: Internet of Things for Smart Cities. ago. 2015, [S.l.]: IEEE, ago. 2015. p. 309–316.

CAMPOS, Guilherme Oliveira. Estudo, avaliação e comparação de técnicas de detecção não supervisionada de outliers. p. 67, 2015.

CARMELO, Monte. Leandro Magno Gomides de Sousa Modelagem e compensação de erro de sensores e atuadores baseados em Arduino Leandro Magno Gomides de Sousa. 2018.

CASATI, Fabio *et al*. Towards business processes orchestrating the physical enterprise with wireless sensor networks BT - 34th International Conference on Software Engineering, ICSE 2012, June 2, 2012 - June 9, 2012. p. 1357–1360, 2012.

DE FRANÇA, Tiago C *et al*. Web das Coisas: Conectando Dispositivos Físicos ao Mundo Digital. **Livro Texto de Minicursos - SBRC 2011**, p. 48, 2011.

FAGUNDES, Priscila Basto. Uma Análise Das Relações Entre a Qualidade Da Informação E Big Data. **XVIII Encontro Nacional De Pesquisa Em Ciência Da Informação – Enancib 2017Encontro Nacional De Pesquisa Em Ciência Da Informação – Enancib 2017**, n. 2014, p. 206–220, 2017.

FAWZY, Asmaa; MOKHTAR, Hoda M O; HEGAZY, Osman. Outliers Detection and Classification in Wireless Sensor Networks. **Egyptian Informatics Journal**, v. 14, n. 2, p. 157–164, 2013.

GOLEMAN, DANIEL; BOYATZIS, RICHARD; MCKEE, Annie. Identificação de Outliers. **Journal of Chemical Information and Modeling**, v. 53, n. 9, p. 1689–1699, 2019.

GRUBBS, Frank E. Procedures for

Detecting Outlying Observations in Samples. **Technometrics**, v. 11, n. 1, p. 1–21, 1969.

GUBBI, Jayavardhana *et al*. Internet of Things (IoT): A vision, architectural elements, and future directions. **Future Generation Computer Systems**, v. 29, n. 7, p. 1645–1660, set. 2013.

HUI, Jonathan; CULLER, David; CHAKRABARTI, Samita. 6LoWPAN Network Architecture. **Architecture**, p. 1–17, 2009.

KAHN, Beverly K.; STRONG, Diane M.; WANG, Richard Y. Information Quality Benchmarks: Product and Service Performance. **Communications of the ACM**, v. 45, n. 4, p. 184–192, 2002.

KEFALAKIS, Nikos *et al*. D4.3.1 Core OpenIoT Middleware Platform. 2013.

KO, JeongGil *et al*. Evaluating the Performance of RPL and 6LoWPAN in TinyOS. **Ip+SN 2011**, 2011.

LEE, Yang W. *et al*. AIMQ: A methodology for information quality assessment. **Information and Management**, v. 40, n. 2, p. 133–146, 2002.

LEZCANO, Leonardo; SANTOS, Leopoldo; GARCÍA-BARRIOCANAL, Elena. Semantic integration of sensor data and disaster management systems: The emergency archetype approach. **International Journal of Distributed Sensor Networks**, v. 2013, 2013.

LIMA, MARLOS ANTONIO DOS SANTOS. **Sistema para Análise de Dados de Nodos Sensores**. 2014. 52 f. Universidade do Estado do Rio Grande do Norte, 2014.

LUCATO, Melissa U; COUTO, Paulo R G; LUZ, Denise. Proposta para o estabelecimento da confiabilidade metrológica em calibração volumétrica. **V Congresso Latino Americano de Metrologia**, v. 2, n. 1, p. 2–5, 2007.

MEYER, Sonja; RUPPEN, Andreas; MAGERKURTH, Carsten. Internet of Things-Aware Process Modeling: Integrating IoT Devices as Business Process Resources. **Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)**. [S.l: s.n.], 2013. v. 7908 LNCS. p. 84–98.

NI, Kevin *et al*. Sensor network data fault types. **ACM Transactions on Sensor Networks**, v. 5, n. 3, p. 1–29, 1 maio 2009.

NISHA, U.BARAKKATH *et al*. Statistical Based Outlier Detection in Data Aggregation for Wireless Sensor Networks. **Journal of Theoretical and Applied Information Technology**, v. 59, n. 3, p. 770–780, 2014.

OLIVEIRA, Elcio Cruz De. Comparação das diferentes técnicas para a exclusão de “outliers”. **Metrologia**, 2008.

OUSSOUS, Ahmed *et al*. Big Data technologies : A survey. **Journal of King Saud University - Computer and Information Sciences**, v. 30, n. 4, p. 431–448, 2018.

SACHIDANANDA, Vinay *et al*. Quality of Information in Wireless Sensor Networks : A Survey. **Proceedings of the 15th International Conference on Information Quality (ICIQ-2010)**, v. 1, p. 1–15, 2010.

SHAIK, Arshad; ESWARAN, P. Removal of IEEE 802 . 15 . 4 MAC Unrelaibility Problem in Hardware Superframe structure. **International Journal of Advanced Computer Research**, v. 2, n. 4, 2012.

YANG ZHANG; MERATNIA, Nirvana; HAVINGA, Paul. Outlier Detection Techniques for Wireless Sensor Networks: A Survey. **IEEE Communications Surveys & Tutorials**, v. 12, n. 2, p. 159–170, 2010.