**UNIVERSIDADE FEDERAL DE SANTA CATARINA**
**DEPARTAMENTO DE AUTOMAÇÃO E SISTEMAS**

Patrícia Bordignon André

# DETECÇÃO E IDENTIFICAÇÃO DE *OUTLIERS* EM REDES DE SENSORES SEM FIO DE LARGA ESCALA

Florianópolis

2017

Patrícia Bordignon André

# DETECÇÃO E IDENTIFICAÇÃO DE *OUTLIERS* EM REDES DE SENSORES SEM FIO DE LARGA ESCALA

Dissertação submetida ao Programa de Pós-Graduação em Engenharia de Automação e Sistemas para a obtenção do Grau de Mestre em Engenharia de Automação e Sistemas.
Orientador: Prof. Carlos Barros Montez, Dr.
Coorientadores: Prof. Ricardo Moraes, Dr. e Prof. Alex Pinto, Dr.

Florianópolis

2017

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

André, Patrícia Bordignon
Detecção e Identificação de Outliers em Redes de Sensores sem Fio de Larga Escala / Patrícia Bordignon André ; orientador, Carlos Barros Montez, coorientador, Ricardo Moraes, coorientador, Alex Pinto, 2017.
117 p.

Dissertação (mestrado) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós Graduação em Engenharia de Automação e Sistemas, Florianópolis, 2017.

Inclui referências.

1. Engenharia de Automação e Sistemas. 2. Redes de Sensores sem Fio. 3. Detecção de Outliers. 4. Identificação de Outliers. 5. Larga Escala. I. Montez, Carlos Barros . II. Moraes, Ricardo . III. Pinto, Alex IV. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Engenharia de Automação e Sistemas. V. Título.

Patrícia Bordignon André

**DETECÇÃO E IDENTIFICAÇÃO DE *OUTLIERS* EM REDES DE SENSORES SEM FIO DE LARGA ESCALA**

Esta Dissertação foi julgada aprovada para a obtenção do Título de "Mestre em Engenharia de Automação e Sistemas", e aprovada em sua forma final pelo Programa de Pós-Graduação em Engenharia de Automação e Sistemas.

Florianópolis, 31 de julho 2017.

---

Prof. Daniel Coutinho, Dr.
Coordenador do PPGEAS
Universidade Federal de Santa Catarina

**Banca Examinadora:**

Prof. Carlos Barros Montez, Dr
Universidade Federal de Santa Catarina
Orientador

Prof. Ricardo Moraes, Dr
Universidade Federal de Santa Catarina
Coorientador

Prof. Alex Pinto, Dr
Universidade Federal de Santa Catarina
Coorientador

Prof. Werner Kraus Junior, Dr
Universidade Federal de Santa Catarina

Profa. Patrícia Plentz, Dra
Universidade Federal de Santa Catarina

Prof. Ubirajara Franco Moreno, Dr
Universidade Federal de Santa Catarina

Este trabalho é dedicado à minha querida Mamadi.

## AGRADECIMENTOS

Sou grata por ter tido a oportunidade de desenvolver este trabalho. Agradeço a Deus por ter me dado forças para seguir em frente em busca do meu sonho.

Agradeço à minha família e especialmente à minha mãe por acreditar no meu sonho e me ajudar a torna-lo realidade.

Agradeço aos professores Carlos Montez, Ricardo Moraes e Alex Pinto pela orientação realizada, o suporte oferecido e a disponibilidade de tempo em sanar minhas dúvidas.

Agradeço aos professores Tadeu e Gustavo pelas contribuições e ajuda oferecida para a conclusão deste trabalho.

Agradeço ao Dani pelos conselhos e ajuda oferecida nos momentos de dificuldade.

Agradeço aos novos amigos que Florianópolis me deu, especialmente Gabriela Torres, por me apoiar e incentivar. E agradeço aos velhos amigos, principalmente Ivana, por todas as conversas e motivações e a Suelen por estar presente desde os tempos do colégio.

Ao decorrer desta jornada os colegas de mestrado tornaram-se amigos. Em especial agradeço aos colegas do LTIC, Jéssica, Renan e Lucas por dividirmos nossas experiências e infinitos momentos de alegria.

Agradeço a todos de que alguma forma contribuiu para o desenvolvido deste trabalho e a todos que me apoiaram e entenderam meus momentos de reclusão social.

Por fim, agradeço a CAPES pelo apoio financeiro concedido.

Torture numbers, and they'll confess to anything.

(Gregg Easterbrook)

## RESUMO

Redes de Sensores Sem Fio (RSSFs) são utilizadas em diversas áreas para rastreamento e monitoramento de ambientes. A facilidade de implantação dessas redes, associada ao baixo custo dos nodos, incentivam a sua utilização para fins comerciais, militares e industriais. Entretanto, as RSSFs de larga escala, por possuírem uma grande quantidade de nodos implantados, geram uma grande quantidade de dados brutos. Além disso, em virtude da própria natureza dessas redes, dados anômalos (*outliers*) podem ser gerados, comprometendo a confiabilidade dos dados. De forma geral, a utilização de técnicas para detecção e identificação (classificação) de *outliers* é essencial para manter a confiabilidade dos dados para que futuras tomadas de decisões sejam realizadas. Devido às restrições de *hardware* dos nodos, as técnicas tradicionais de detecção e identificação de *outliers* geralmente não são aplicáveis às RSSFs. Sendo assim, a aplicação de técnicas de baixo custo computacional é uma das únicas soluções viáveis. O objetivo desse trabalho de mestrado é analisar e aplicar técnicas de detecção e identificação de *outliers*, de baixo custo computacional, para RSSFs de larga escala. A abordagem proposta é dividida em duas etapas: a primeira para detecção de *outliers*, através da aplicação de técnicas baseadas em estatísticas. A segunda etapa é dedicada à identificação de *outliers*, por meio da combinação de correlações espaciais e limites pré-definidos. Para avaliação da proposta é utilizado o simulador OMNeT++/Castalia. Os resultados obtidos através das simulações mostraram que é viável a utilização de técnicas baseadas em estatísticas com baixo custo computacional, para a detecção e identificação de *outliers* em RSSF de larga escala.



**Palavras-chave:** Redes de Sensores Sem Fio. Detecção de *Outlier*. Identificação de *Outlier*. Larga Escala.

## ABSTRACT

Wireless Sensor Networks (WSNs) are used in many areas for tracking and monitoring environments. The ease of deployment of these networks, coupled with the low cost of nodes, encourages their use for commercial, military and industrial purposes. However, the large-scale WSNs have a large number of implanted nodes, consequently generate a large amount of raw data. Moreover, due to the nature of these networks, outliers can be generated, compromising data reliability. In general, the use of techniques for detection and identification (classification) of outliers to maintain the reliability of data for future detainees is vital. Due to the hardware constraints of the nodes, traditional outliers detection and identification techniques are generally applicable to WSNs. Thus, an application of techniques of low computational cost is one of the only solution viable. The objective of this masters dissertation is to analyze and apply low-computational outliers detection and identification techniques for large-scale WSNs. The proposed approach is divided into two steps: the first one to detect outliers, through the application of statistical techniques. The second step is dedicated to the identification of outliers, through the combination of spatial correlations and predefined boundaries. To evaluate the proposal, the OMNeT ++/Castalia simulator is used. The results obtained through the simulations showed that it is feasible to use techniques based on statistics with low computational cost for the detection and identification of outliers in large-scale WSN.



**Keywords:** Wireless Sensor Network. Outlier Detection. Identification Outlier. Large-Scale.

## LISTA DE FIGURAS

## LISTA DE TABELAS

## LISTA DE ABREVIATURAS E SIGLAS

| | |
|---|---|
| ADC | Analog-to-Digital |
| AOD | Adaptive Outlier Detection |
| APCCAD | Adaptive Principal Component Classifier-based Anomaly Detection |
| CCA | Clear Channel Assessment |
| FTA | Fault Tolerant Averaging |
| FTWOD | Fixed-size Time Window-based Outlier Detection |
| GPS | Global Positioning System |
| GTS | Guaranteed Time Slots |
| IOD | Instant Outlier Detection |
| IoT | Internet of Things |
| ISM | Industrial Sientific and Medical |
| KDE | Kernel Density Estimator |
| LQI | Link Quality Indication |
| LR-WPAN | Low-Rate Wireless Personal Area Networks |
| MAC | Media Access Control |
| MAD | Median Absolute Deviation |
| MCTI | Ministério da Ciência, Tecnologia e Inovação |
| MEMS | Micro Electro-Mecanical Systems |
| OOD | Online Outlier Detection |
| PCA | Principal Component Analysis |
| PCCAD | Principal Component Classifier-based Anomaly Detection |
| PHY | Physical Layer |
| PPDU | PHY Protocol Data Unit |
| QS-SVM | Quarter Sphere Support Vector Machine |
| ROC | Receiver-Operating Characteristic |
| RSL | Revisão Sistemática de Literatura |
| RSSF | Rede de Sensores Sem Fio |
| RSSI | Received Signal Strength Indicator |
| SOSUS | Sound Surveillance System |
| SVM | Support Vector Machines |
| WPAN | Wireless Personal Area Networks |

WSN Wireless Sensor Networks

## SUMÁRIO

# 1 INTRODUÇÃO

O uso das Redes de Sensores Sem Fio (RSSFs) vem crescendo cada vez mais com o desenvolvimento de tecnologias para sensores inteligentes ("*smart sensors*") (YICK; MUKHERJEE; GHOSAL, 2008). Essas redes são compostas por dezenas, centenas ou milhares de nodos que monitoram e controlam grandezas físicas de um determinado ambiente (AKYILDIZ et al., 2002a). O escopo de aplicações dessas redes vem se ampliando, abrangendo aplicações de sensoriamento em diversas áreas, como: monitoramento de áreas militares, monitoramento remoto de pacientes, monitoramento de áreas de difícil acesso ou de risco, monitoramento ambiental, rastreamento de veículos, automação residencial e industrial, entre muitas outras (AKYILDIZ et al., 2002a; ZHANG; MERATNIA; HAVINGA, 2010).

Em geral, os nodos possuem capacidades de sensoriamento, processamento, memória e comunicação de dados. Apesar de apresentarem restrições de recursos eles são utilizados em larga escala, devido aos seus baixos custos de aquisição (AKYILDIZ et al., 2002a; ILYAS; MAHGOUB, 2004). Esses nodos geralmente são distribuídos por toda a área a ser monitorada, de modo que possam coletar os dados do ambiente e encaminhar a uma estação base. Como consequência da grande quantidade de nodo utilizados, as RSSFs produzem grandes volumes de dados brutos, que precisam ser processados, com a finalidade de minimizar o tráfego da rede (NAKAMURA; LOUREIRO; FRERY, 2007).

A norma IEEE 802.15.4 surgiu com objetivo de padronizar a comunicação entre dispositivos/nodos de baixa potência e para prover a interoperabilidade entre eles. Nela, a camada física e de controle de acesso ao meio foram especificadas. Neste cenário de interoperabilidade surgiu mais recentemente o conceito da Internet das Coisas, a qual deverá coexistir com as RSSFs, integrando os objetos do mundo físico em uma infraestrutura de comunicação global (PANTELAKI; PANAGIOTAKIS; VLISSIDIS, 2016).

Devido ao fato dos nodos apresentarem recursos computacionais reduzidos e/ou o ambiente apresentar características hostis, as leituras dos sensores podem conter dados anômalos (*outliers*) e não serem consideradas confiáveis (LOUREIRO et al., 2003). Neste cenário de desenvolvimento com grandes volumes de dados, surge a necessidade de mantê-los confiáveis e livres desses dados anômalos para que as tomadas de decisões possam ser executadas corretamente. Portanto, a utilização das técnicas para detecção e identificação (classificação) de *outliers* são

necessárias (NAKAMURA; LOUREIRO; FRERY, 2007).

Dados anômalos ou *outliers* são aqueles dados que desviam significantemente do conjunto padrão de leituras, ou apresentam alguma inconsistência quando comparados com os demais dados (HAWKINS, 1982; SHENG et al., 2007; BARNETT; LEWIS, 1994). Nas RSSFs, as anomalias podem ser classificadas em três categorias: anomalia no nodo, anomalia da rede e anomalia nos dados. Este trabalho está focado na detecção de anomalias nos dados. Essas anomalias ocorrem quando um valor sensoriado apresenta discrepâncias temporal, espacial ou espaço-temporal em relação aos demais valores.

Os dados anômalos por sua vez podem ter origem em três diferentes fontes: (i) ruídos e erros (ex. um sensor defeituoso ou mal calibrado), (ii) eventos (ex. incêndios florestais) ou (iii) ataques maliciosos (Negação de Serviço - DoS) (ZHANG; MERATNIA; HAVINGA, 2010).

A detecção de *outliers*[1] nos dados é motivada, principalmente, por assegurar a confiabilidade, e garantir a robustez dos dados analisados. O processo de identificação é a etapa posterior que determina se o *outlier* detectado (podendo ser um ou mais que um) é resultante de um evento relevante ou é um dado espúrio decorrente, por exemplo, de um sensor defeituoso ou com ruído. Importante destacar que os processos de detecção e identificação precisaram ser capazes de se adaptar à dinamicidade intrínseca das RSSFs (ZHANG; MERATNIA; HAVINGA, 2010; BHOJANNAWAR; BULLA; DANAWADE, 2013).

Os processos de detecção das anomalias podem ocorrer de duas formas: *online* e *offline*. No modo *online*, o *outlier* é detectado logo após a sua leitura. Por outro lado, no modo *offline* a detecção é feita após os dados serem enviados para uma estação base (BHOJANNAWAR; BULLA; DANAWADE, 2013).

As técnicas de detecção de *outliers* tradicionais geralmente não são aplicáveis às RSSFs, devido as restrições de recursos das RSSFs (ZHANG; MERATNIA; HAVINGA, 2010). Para tal, novas abordagens devem ser analisadas.

No contexto do que foi apresentado, este trabalho busca melhorar os processo de detecção e identificação de *outliers*. Neste sentido, busca-se não apenas a detecção do dado anômalo, mas também a sua identificação para determinar se este é resultante de um evento relevante ou é apenas um dado espúrio, para um posterior correto tratamento deste *outlier*.

A abordagem proposta é dividida em duas etapas: a primeira,

[1]Neste texto os termos "*outliers*" e "dados anômalos" serão usados de forma intercambiável.

para detecção dos *outliers*, utiliza técnicas de detecção baseadas em estatísticas, devido ao seus baixos custos computacionais. A segunda etapa, identificação dos *outliers*, utiliza a combinação de duas técnicas: correlações baseadas em informações obtidas com os nodos vizinhos e limites pré-definidos.

Para o levantamento do estado da arte e trabalhos correlatos optou-se por realizar uma Revisão Sistemática da Literatura. A partir desses trabalhos foram identificadas questões poucos abordadas, como identificação de eventos em RSSFs de larga escala, são nessas questões que este trabalho irá se aprofundar.

Por fim, o problema de pesquisa que guia este trabalho pode ser resumido na seguinte questão: "É possível utilizar técnicas com baixo custo computacional que permitam detectar e identificar *outliers* em redes de sensores sem fio de larga escala?"

## 1.1 OBJETIVO

O principal objetivo deste trabalho é analisar e aplicar técnicas de baixo custo computacional que permitam detectar e identificar *outliers* em RSSFs de larga escala.

### 1.1.1 Objetivos Específicos

Para alcançar o objetivo geral proposto neste trabalho, os seguintes objetivos específicos surgem:

- Analisar e aplicar técnicas baseadas em estatísticas que permitam detectar *outliers*, e propor ou combinar técnicas que permitam identificar os *outliers* detectados;
- Definir métricas de desempenho para avaliar as técnicas;
- Modelar um cenário de simulação de RSSFs de larga escala;
- Avaliar e validar a abordagem proposta para a detecção e identificação de *outliers* em RSSFs no ambiente proposto.

## 1.2 MÉTODO DE PESQUISA

Para a execução dessa dissertação foi realizada uma pesquisa exploratória e tecnológica. Para atingir os resultados os seguintes pro-

cedimentos foram realizados.

- Uso de um método de Revisão Sistemática da literatura, para identificar o estado da arte sobre detecção e identificação de *outliers* em RSSF;
- Proposição de uma abordagem para detectar e identificar *outliers* com base nos resultados obtidos por meio da Revisão Sistemática da Literatura;
- Utilização de uma ferramenta de simulação de RSSFs OMNeT++/Castalia para avaliar a proposta.

## 1.3 LIMITAÇÕES DO ESCOPO DO TRABALHO

A fim de delimitar a abrangência do trabalho, algumas questões não são discutidas. As anomalias de nodo e rede não são consideradas neste trabalho, bem como, as anomalias de dados causadas por ataques maliciosos. Este trabalho não busca avaliar protocolos adequados para a comunicação entre os nodos. Assume-se que todos os nodos possuem comunicação direta ou por *multihop* com o coordenador. As questões energéticas não são diretamente tratadas. Contudo, são utilizadas técnicas de baixo custo computacional e que não envolvem troca de dados entre os nodos. Uma hipótese assumida neste trabalho é que as técnicas propostas são adequadas para uso em RSSF de larga escala.

## 1.4 ORGANIZAÇÃO DO TEXTO

Esta dissertação está organizada em seis capítulos. No primeiro capítulo foi apresentada a introdução, os objetivos, a metodologia utilizada, e as limitações do escopo de abrangência deste trabalho. O segundo capítulo é dedicado à fundamentação teórica das principais áreas de pesquisa abordadas por este trabalho como RSSF, técnicas de detecção e identificação de *outliers* e os conceitos teóricos necessários para o entendimento e embasamento para este trabalho. O terceiro capítulo possui os trabalhos relacionados encontrados através da Revisão Sistemática da Literatura, mostrando as principais lacunas a serem exploradas. Em seguida, no quarto capítulo é descrita a proposta. No quinto capítulo é feita a avaliação da proposta. Por último, no sexto capítulo, os resultados obtidos são discutidos e são tecidas as considerações finais.

# 2 FUNDAMENTAÇÃO TEÓRICA

Os avanços tecnológicos impulsionados pelos semicondutores possibilitaram o desenvolvimento e aprimoramentos dos MEMS (*Micro Electro-Mechanical Systems*) e dos sensores inteligentes. A expectativa da produção desses sensores é que atinja grandes quantidade para reduzir os custos e difundir a tecnologia junto com as Redes de Sensores Sem Fio (RSSFs) (LOUREIRO et al., 2003).

Segundo o Silicon Laboratories (2013), muitas dessas inovações tecnológicas surgiram motivadas por aplicações militares. A primeira rede sem fio considerada como uma RSSF foi a *Sound Surveillance System* (SOSUS). Ela foi desenvolvida pelos militares norte americanos para detectar e rastrear submarinos soviéticos na década de 1950. Essa rede foi implantada nos oceanos Atlântico e Pacíficos com sensores acústicos. Hoje em dia a rede ainda está ativa, não mais com fins militares, mas sim para monitoramento da vida marinha e atividades vulcânicas (Silicon Laboratories, 2013).

Neste capítulo apresentam-se conceitos e definições relacionados às RSSFs e tratamento de *outliers* e uma revisão bibliográfica. O capítulo está dividido em três seções. A primeira visa apresentar as RSSFs e suas aplicações, os nodos, Internet das Coisas, Sensores e o padrão IEEE 802.15.4. A segunda seção é dedicada aos conceitos referentes à detecção e identificação de *outliers*, apresentando as classificações das técnicas. Por fim, na terceira seção as considerações do capítulo são discutidas.

## 2.1 REDES DE SENSORES SEM FIO

Redes de Sensores sem Fio, são geralmente, compostas por um grande número de nodos. Esses nodos podem estar posicionados dentro do fenômeno observado ou próximo a ele (AKYILDIZ et al., 2002a). Segundo Ilyas e Mahgoub (2004), os nodos geralmente são pequenos e podem possuir características limitadas de sensoriamento, potência e processamento.

Uma RSSF é projetada para detectar eventos ou fenômenos, coletar e processar dados, e transmiti-los para uma estação base ou nodo *sink*[1], normalmente por um canal de rádio frequência. O nodo *sink* geralmente é um tipo especial de nodo, pois ele possui maior autono-

[1] Neste trabalho estação base e *sink* serão utilizados como sinônimos.

mia em recursos energéticos e alcance de rádio, entretanto não realiza sensoriamento (RUIZ et al., 2004).

Na bibliografia estudada não encontrou-se um consenso ou uma definição exata de quantos nodos são necessários para denominar uma rede como sendo de larga escala. De acordo com o Chouikhi et al. (2015), a quantidade de nodos caracterizar se a rede é de larga escala ou não. Uma rede com dezenas de nodos é considerada de pequena escala, enquanto que uma rede com centenas ou milhares de nodos, é de larga escala (CHOUIKHI et al., 2015). Entretanto, os trabalhos de Zhang et al. (2004) e Cerpa et al. (2001) utilizam RSSFs que possuem 50 e 150 nodos respectivamente e são considerados de larga escala. Outro ponto importante, é que a rede deve ser escalável para que seja considerada de larga escala (WANG et al., 2014).

Apesar das limitações encontradas nas RSSFs, o seu uso vem se tornando cada vez mais frequente. Estão presentes nas mais diversas áreas de atuação como: controle (monitoramento de ferramenta na manufatura), monitoramento ambiental (desastres naturais e incêndios), controle de tráfego de veículos em rodovias, segurança, ambientes hospitalares para monitoramento de pacientes, e no uso militar para a identificação de ameaças entre outras aplicações (AKYILDIZ et al., 2002b; ILYAS; MAHGOUB, 2004; LOUREIRO et al., 2003).

Segundo Yick, Mukherjee e Ghosal (2008) podemos dividir as aplicações em RSSFs em duas categorias: rastreamento e monitoramento. Aplicações de monitoramento abrangem tanto ambientes internos quanto externos; enquanto que as aplicações de rastreamento incluem rastreamento de animais, pessoas, objetos e veículos. Cada categoria pode ser divida em várias subcategorias:

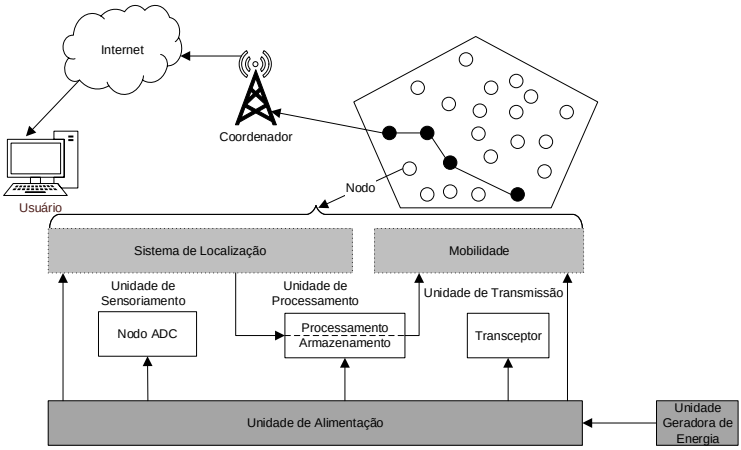
- **Rastreamento**:
  - Militar: rastreamento de inimigos;
  - *Habitat*: rastreamento de animais;
  - Negócios: rastreamento de pessoas;
  - Público/Industrial: rastreamento de tráfego e veículos.
- **Monitoramento**:
  - Militar: detecção de segurança;
  - *Habitat*: monitoramento de animais;
  - Negócios: sistemas de gestão;
  - Público/Industrial: monitoramento estrutural, industrial, gestão, maquinas pesadas, químico;

– Saúde: monitoramento de pacientes;
– Ambiental: monitoramento ambiental (clima, temperatura, pressão).

### 2.1.1 Estrutura do Nodo

Uma RSSF é formada por nodos, os quais são dispositivos autônomos equipados com quatro unidades básicas: unidade de sensoriamento (geralmente analógica), unidade de processamento (processador e memória), unidade de comunicação (emissor e receptor) e unidade de alimentação/energia (AKYILDIZ et al., 2002b). A Figura 1 ilustra uma típica RSSF com ênfase nas unidades de um nodo.

Figura 1 – Estrutura de um nodo.




Fonte: Rassam, Zainal e Maarof (2013).

A unidade de sensoriamento é responsável por capturar as mudanças físicas que ocorrem no fenômeno monitorado, como temperatura e pressão, e transformar as grandezas físicas em sinais digitais (através de um conversor analógico digital ADC) para ser processado pela unidade de processamento (RASSAM; ZAINAL; MAAROF, 2013).

A unidade de processamento é responsável por processar as grandezas capturadas dos sensores, e armazenar as informações necessárias para o processamento das tarefas. Já a unidade de Transmissão propor-

ciona um meio para transferir os sinais para uma estação base, outros nodos, ou uma rede de computadores (VIEIRA et al., 2003).

Unidades adicionais podem ser incorporadas ao nodo, como o GPS (*global positioning system*) com o objetivo de localizar a posição de nodos móveis. Os nodos também podem ter mobilidade, dependendo da aplicação. Por fim, os nodos também podem ter uma unidade geradora de energia, como por exemplo, placas solares (AKYILDIZ et al., 2002b).

Os nodos muitas vezes são posicionados em um ambiente hostil estando suscetíveis as interferências eletromagnéticas, à influência do meio ambiente, e a falhas de *hardware*. As posições dos nodos podem ou não ser conhecidas, entretanto, os protocolos de rede e algoritmos precisam ter capacidade de auto-organização e precisam ser adaptativos, para lidar com a dinamicidade das RSSFs(AKYILDIZ et al., 2002b).

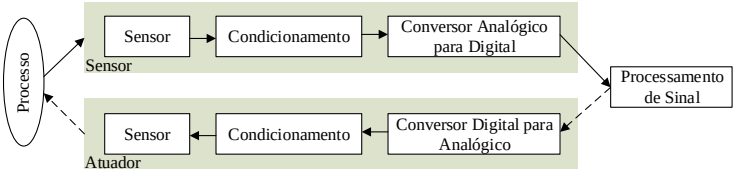
### 2.1.2 Sensoriamento e Sensores

Segundo Dargie e Poellabauer (2010), *sensoriamento* é uma técnica usada para coletar informações sobre um processo físico ou objeto, bem como eventos. O objeto que realiza esta tarefa é chamado de sensor ou transdutor. O sensor é um dispositivo que converte parâmetros ou eventos do mundo físico em sinais que podem ser medidos e analisados. É comum também encontrar atuadores em RSSF, os quais são dispositivos que podem atuar/modificar o processo físico.

Na Figura 2 o processo de aquisição das informações e atuação no meio físico é apresentado. O sensor realiza uma leitura de um processo físico. Como resultado dessa leitura, temos um sinal elétrico que precisa ser condicionado, o qual é um processo de aplicação de filtros para a remoção de ruídos. Após a aplicação dos filtros, o sinal está pronto para ser convertido de analógico para digital. Por fim, este sinal estará disponível para ser processado, armazenado ou visualizado (DARGIE; POELLABAUER, 2010; CALLEGARO, 2014).

A escolha do sensor está relacionada com o tipo de aplicação e da propriedade física a ser monitorada, por exemplo, temperatura, pressão, luminosidade entre outros. Os sensores podem ser classificados com base em diversos métodos. Na Tabela 1, a classificação foi feita com base nas características físicas comuns entre eles. Outra classificação possível é através do fenômeno elétrico usado para converter as propriedades físicas em sinais elétricos, por exemplo: resistivo, capacitivo, indutivo, campo magnético (ILYAS; MAHGOUB, 2004; DARGIE;

Figura 2 – Processo de sensoriamento.




Fonte: Adaptado de Dargie e Poellabauer (2010).

POELLABAUER, 2010).

Tabela 1 – Classificação dos sensores.

| **Tipo** | **Exemplos** |
|---|---|
| Temperatura | Termistores, termopares. |
| Pressão | Manômetros, barômetros, medidores de ionização. |
| Ótico | Fotodiodos, foto transístores e infravermelho. |
| Acústico | Ressonadores piezoeléctricos e microfones. |
| Mecânico | Medidores de tensão, sensores táteis capacitivos, diafragmas e células piezoresitivos. |
| Movimento/ Vibração | Acelerômetros, giroscópios e sensores fotográficos. |
| Fluxo | Anemômetro e fluxo de ar. |
| Posição | GPS, sensores de ultrassom, sensores infravermelhos e inclinômetros. |
| Eletromagnetismo | Sensores de efeito Hall e magnetômetros. |
| Químico | Sensores de ph, sensores eletroquímicos e sensores de gás. |
| Umidade | Sensores capacitivos e resistivos e higrômetro. |
| Radiação | Detectores de ionização e Geiger-contadores Mueller. |

Fonte: Dargie e Poellabauer (2010).

Existem também sensores comumente utilizados na área médica. O objetivo destes é monitorar os sinais vitais dos pacientes. Alguns exemplos são: pressão sanguínea, eletrocardiograma, eletromiograma, temperatura, oxigenação, pulso entre outros.

### 2.1.3 Internet das Coisas

A partir das RSSFs, um novo paradigma computacional vem se tornando realidade, denominado de Internet das Coisas ou do inglês *Internet of Things* (IoT). Esse conceito propõe que cada objeto ou "coisa" pode ser equipado com sensores que se comunicam entre si, formando uma rede de cooperação com o intuito de realizar tarefas em comum (ATZORI; IERA; MORABITO, 2010; GIUSTO et al., 2010; RUIZ et al., 2016).

Segundo Bassi e Horn (2008), a IoT, deverá ser uma rede mundial de objetos interconectados por uma identificação única, baseados em protocolos de comunicação 6LoWPAN(*IPv6 over Low power Wireless Personal Area Network*). Segundo Miorandi et al. (2012), a Internet das coisas possui características que precisam ser suportadas pela rede, tais como:

- **Heterogeneidade**: Principal característica da IoT, devido à grande diversidade e heterogeneidade dos dispositivos que fazem parte da rede. Cada dispositivo possui diferentes capacidades computacionais e de comunicação.
- **Escalabilidade**: A questão sobre escalabilidade deve ser tratada em diferentes níveis para suportar a inclusão de novos objetos: (a) nomeação e endereçamento, (b) comunicação de dados, (c) gerenciamento da informação e conhecimento e (d) gerenciamento de serviços.
- **Onipresença**: A comunicação de dados precisa acontecer de maneira onipresente através das redes sem fio.
- **Otimização de recursos energéticos**: A questão das baterias dos dispositivos geram restrições nas tarefas de comunicação e computação. Sendo assim, soluções que otimizem o uso de energia tornam-se importantes para a IoT.
- **Localização e rastreamento**: Os objetos que fazem parte da IoT, estando em movimento ou não, podem necessitar ser rastreados.
- **Auto-organização**: Devido à dinâmica da IoT, os dispositivos podem estar associados ou não à rede. Utilizar mecanismos de auto-organização permite que os nodos se organizem de forma autônoma.

- **Interoperabilidade e gerenciamento de dados**: Devido à grande heterogeneidade dos dispositivos, garantir a interoperabilidade é um desafio a ser conquistado. Para tornar útil a grande quantidade de dados gerados pela IoT, é necessário fornecer dados com formatos adequados e padronizados, ou seja, utilizar uma descrição semântica para os metadados e utilizar linguagens e formatos bem definidos.

- **Segurança e privacidade**: Mecanismos que assegurem a segurança e privacidade dos dados é fundamental para garantir a aceitação por parte dos usuários. Esses mecanismos devem ser concebidos na própria arquitetura das soluções IoT.

Além dessas características que precisam ser suportadas, segundo Oliveira (2011) e Gubbi et al. (2013), a IoT possui outros sérios desafios, tais como: gerenciamento remoto, usabilidade, privacidade e qualidade de serviço.

Um dos efeitos causados pela IoT é a grande produção de dados brutos. No contexto das RSSF aplicadas à IoT, um dos principais desafios é o de analisar e extrair esses dados em busca de informações relevantes. Neste sentido, a detecção e identificação de *outliers* destaca-se como uma importante tarefa a ser cumprida por garantir a confiabilidade dos dados.

No âmbito da IoT e, mais especificamente, das RSSF, a partir dos esforços para padronizar uma rede de comunicação que oferecesse a interoperabilidade necessária e com as restrições dos nodos levadas em consideração, surgiu o protocolo padronizado como IEEE 802.15.4.

### 2.1.4 IEEE 802.15.4

As RSSF, quando comparadas com as redes de computadores tradicionais, apresentam necessidades e características diferentes. Neste sentido, o protocolo IEEE 802.15.4 é adequado para as redes LR-WPAN (*Low-Rate Wireless Personal Area Networks*) (CALLEGARO, 2014).

O protocolo IEEE 802.15.4 é considerado como padrão para as camadas físicas (PHY) e de controle de acesso ao meio (MAC) para redes LR-WPAN que possuem taxas de transferências de dados baixas, fontes de energia com restrições e comunicações de curto alcance. O objetivo desse tipo de rede é a facilidade na configuração, transferência de dados, baixo custo e uma vida útil razoável. Tudo isso suportado por um protocolo simples e flexível (IEEE, 2015).

O protocolo IEEE 802.15.4 está fortemente atrelado às tecnologias ZigBee[2], ISA100.11a[3], WirelessHART[4], MiWi[5] e Thread[6]. Cada tecnologia desenvolve as suas camadas superiores não especificadas pelo padrão IEEE 802.15.4.

Existem dois tipos de dispositivos para o protocolo IEEE 802.15.4: os de função completa FFD (*full function devices*) e os de função reduzida RFD (*reduced function device*). Os FFD podem ser utilizados como coordenador de uma rede ou cliente. Os dispositivos RFD são utilizados em aplicações mais simples, devido ao seu limite de recursos e não podem ser utilizados como coordenador. Se a rede for composta por apenas dispositivos RFD, é recomendável que o coordenador seja um FFD (CALLEGARO, 2014).

As topologias suportadas pelo padrão são estrela (*single-hop*), *cluster-tree* e *mesh* (*multi-hop*) (GUGLIELMO; BRIENZA; ANASTASI, 2016; PANTELAKI; PANAGIOTAKIS; VLISSIDIS, 2016; IEEE, 2015):

- **Estrela**: topologia mais simples, na qual todos os dispositivos estão conectados com o coordenador. O coordenador é responsável pelo controle do tráfego da rede. A maior vantagem desta topologia é em relação ao fluxo de mensagens, precisando de, no máximo, dois saltos (*hops*) para entregar a mensagem ao destino.

- ***Cluster***: nesta organização, a rede é composta por redes estrelas conectadas entre si através dos coordenadores de cada rede, chamados de *cluster heads*.

- ***Tree***: nesta organização, o coordenador é a base (raiz) da *tree* enquanto os outros nodos são nodos folhas ou filhos. Geralmente, se adota uma disciplina de comunicação denominada *convergecast*, onde os nodos filhos não podem conectar-se entre si; eles apenas podem comunicar-se com o coordenador.

- ***Mesh***: topologia *mesh* é formada por um coordenador, dispositivos e vários caminhos. Para uma mensagem chegar ao destino ela passa por vários *hops*. Caso um caminho não esteja funcionando, a rede se auto-organizará para encontrar um caminho alternativo. A maior vantagem desta topologia é a fácil associação e desassociação dos nodos.

[2] www.zigbee.org
[3] www.isa.org
[4] www.hartcomm.org
[5] www.microchip.com
[6] www.threadgroup.org

#### 2.1.4.1 Camada PHY

Segundo o IEEE (2015), as funções da camada física são ativação e desativação do rádio *transceptor*, verificação de energia e sinal (LQI-*link quality indication*), seleção do canal, CCA (*clear channel assessment*), além de permitir a transmissão e recepção das unidade de dados do protocolo (PPDUs). O protocolo opera regularmente nas principais faixas de frequências licenciadas: 868-868.6 MHz (1 canal na Europa), 902-928 MHz (10 canais na América) e 2400-2483.5 MHz (16 canais destinados para ISM (*Industrial Sientific and Medical*) (PANTELAKI; PANAGIOTAKIS; VLISSIDIS, 2016).

#### 2.1.4.2 Subcamada MAC

A subcamada MAC fornece o controle do fluxo de *frames* que passam pela interface do rádio e são transmitidos. Segundo o IEEE (2015), as funções dessa subcamada são gerenciamento dos *beacon*, acesso aos canais, gerenciamento do GTS (*guaranteed time slots*), validação de *frames*, reconhecimento de *frames* recebidos, associação e desassociação dos nodos na rede, além de prover estruturas para implementação de mecanismo de segurança

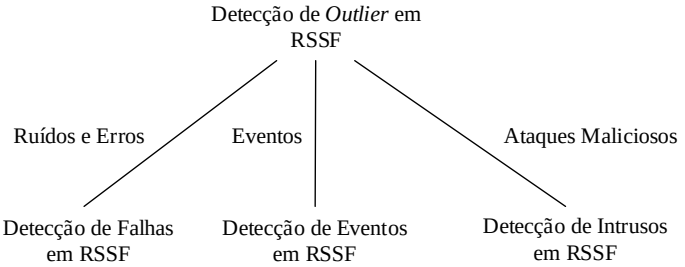
Quando um dispositivo precisa enviar um pacote, essa subcamada solicita para a camada física verificar se o meio está ocupado através do protocolo CSMA/CA. Após a camada física sinalizar que o meio está livre, a subcamada MAC transmite o pacote. Entretanto, se a camada física verificar que o meio está ocupado, a subcamada MAC aguarda durante um tempo aleatório antes de tentar enviar seu pacote novamente (PANTELAKI; PANAGIOTAKIS; VLISSIDIS, 2016; LEÓN; HERNÁNDEZ-SERRANO; SORIANO, 2010).

## 2.2 *OUTLIER*

O termo *outlier* ou anomalia é originário do campo da estatística. Segundo Hawkins (1982), um *outlier* é uma observação que se desvia das demais, criando suspeitas de que esta observação foi gerada por outro mecanismo. Em RSSFs, um *outlier* pode ser assumido como um conjunto de medidas que desviam significativamente ou parecem estar inconsistentes quando comparadas com o restante do conjunto de dados sensoriados (SHENG et al., 2007; BARNETT; LEWIS, 1994).

De acordo com os autores Rassam, Zainal e Maarof (2013), Zhang, Meratnia e Havinga (2010) e Bhojannawar, Bulla e Danawade (2013), os *outliers* podem ser classificados em três fontes: ruídos e erros, eventos e ataques maliciosos, conforme ilustrado na Figura 3.

Figura 3 – Fontes de *outliers* em RSSFs.




Fonte: Zhang, Meratnia e Havinga (2010).

*Outliers* causados por ruídos e erros são aqueles produzidos por nodos defeituosos ou que apresentam alguma inconformidade. Os erros geralmente são mais frequentes de ocorrer, quando comparados com *outliers* identificados como eventos. Erros influenciam a qualidade/precisão dos dados, por este motivo eles devem ser identificados e tratados para não interferir no resultado das possíveis futuras tomadas de decisões (SHAHID; NAQVI; QAISAR, 2015).

Os *outliers* identificados como eventos são alterações nas leituras causados por mudanças físicas do estado no ambiente sensoriado. Por exemplo, incêndio florestal, derramamento de produtos químicos, terremotos entre outros. *Outliers* causados por eventos normalmente afetam vários nodos próximos entre si modificando o padrão das leituras. Por fim, temos *outliers* causados por ataques maliciosos, gerados por intrusos que utilizam de técnicas como, negação de serviço (DoS), *blackhole* entre outros com propósito de contaminar os dados sensoriados (ZHANG; MERATNIA; HAVINGA, 2010; BHOJANNAWAR; BULLA; DANAWADE, 2013).

As RSSFs também estão sujeitas às falhas de *hardware*, de comunicação e ataques maliciosos. As anomalias de RSSFs podem ser classificadas em três categorias (BHOJANNAWAR; BULLA; DANAWADE, 2013; JURDAK et al., 2011):

- **Anomalias no nodo:** são causadas por falhas nos nodos individualmente. Não estão associadas a problemas de comunicação

entre os nodos vizinhos. Geralmente o problema está na bateria que não possui carga o suficiente; ou ainda pode ser decorrente de falha de *hardware* do nodo, como problemas na memória, no rádio, processador, sensores entre outros.

- **Anomalias de rede:** geralmente ocorrem em conjuntos de nodos, e frequentemente estão relacionados com a comunicação entre os nodos da RSSF. As principais causas podem ser: perda da conectividade, *loops* de roteamento, intermitência na conectividade e tempestade de *broadcasting*.

- **Anomalias de dados:** ocorrem quando um valor monitorado apresenta irregularidade quando comparado com os demais valores. Anomalias de dados podem ser facilmente confundidas com anomalias de nodo. Entretanto, só é considerada uma anomalia de nodo quando o problema for produzido por sensores defeituosos. Anomalia de dados pode ser classificada em três tipos:

  - **Temporal:** indica mudanças nas leituras ao longo do tempo. Essas mudanças podem ter vários significados, tais como: (i) grandes alterações nas leituras podem implicar em mudanças do espaço físico, ou seja, indica que um evento foi detectado; e (ii) valores contínuos e fixos durante um longo período de tempo podem indicar que o sensor está defeituoso.
  - **Espacial:** os valores sofrem mudanças em comparação ao nodos próximos (vizinhos). Quando houver um grande desvio dos valores, isso provavelmente indica que o sensor está com defeito ou precisa ser recalibrado.
  - **Espaço-temporal:** os valores combinam alterações com relação aos vizinhos e ao longo do tempo.

Existem dois modos de operação em RSSFs para as técnicas de detecção de *outlier*: *online* e *offline* (RASSAM; ZAINAL; MAAROF, 2013; BHOJANNAWAR; BULLA; DANAWADE, 2013).

- ***Offline***: a detecção feita após um período de tempo é chamada de detecção *offline*, os dados são enviados para uma estação base e processados posteriormente. Esta abordagem implica em um atraso na detecção, ou seja, pode afetar a integridade dos dados. Portanto, este modo de operação não é ótimo para aplicações com restrições de tempo.

- ***Online***: o processo de detecção que ocorre imediatamente após a leitura do sensor é caracterizada como detecção *online*. Este modo de operação possui algumas desvantagens tais como o alto consumo dos recursos da rede: processamento, armazenamento, energia e largura de banda. Para minimizar os custos, as técnicas devem ser de baixo custo. A detecção *online* mantém a integridade dos dados e diminui o tempo para a detecção dos *outliers*.

O tipo dos dados de entrada determina qual técnica de detecção de *outliers* deve ser utilizada para analisar os dados. Dois aspectos costumam ser levados em consideração: atributos e correlações (RASSAM; ZAINAL; MAAROF, 2013; BHOJANNAWAR; BULLA; DANAWADE, 2013; ZHANG; MERATNIA; HAVINGA, 2010; SHAHID; NAQVI; QAISAR, 2015).

- **Atributos**: Os dados podem ser classificados conforme as suas dimensões, em univariados e multivariados. Dados univariados são aqueles lidos por um único tipo de sensor. Representa apenas uma grandeza física, tal como, um sensor responsável por apenas capturar valores de temperatura. Enquanto que dados multivariados são leituras vindas de um único nodo, porém equipado com mais de um tipo de sensor, por exemplo, um nodo que monitora as grandezas físicas como, temperatura, umidade e pressão. Os dados multivariados podem sobrecarregar a rede devido à grande quantidade de dados. Portanto, é preferível escolher técnicas de detecção de *outliers* que utilizem redução de dimensão de dados anômalos com o intuito de prolongar a vida útil da RSSF.

- **Correlações**: Existem os seguintes tipos de dependência entre os dados de cada nodo.

  - **Dependências entre os atributos**: Atributos multivariados podem conter uma relação de dependência ou correlação entre os dados. Por exemplo, as leituras de umidade e pressão podem estar diretamente relacionadas com as leituras de temperatura. É de grande importância determinar essas correlações para melhorar a detecção de erros e eventos.

  - **Dependência das leituras do nodo em seu histórico**: As leituras apresentam correlação temporal. Ocorre quando uma leitura capturada num determinado instante de tempo está correlacionada, com leituras capturadas num instante de tempo anterior. Em ambientes hostis é possível identificar

mudanças frequentes nas distribuições de dados ao longo do tempo.

– **Dependência da leitura do nodo em seus nós vizinhos**: Os nodos posicionados geograficamente próximos apresentam correlações nas suas leituras, chamado de correlação espacial.

A utilização de técnicas que adotem o uso de correlações auxilia a identificar os dados com erros e eventos em RSSFs.

As principais técnicas de detecção de *outliers* em RSSFs são classificadas em três modelos de estrutura (RASSAM; ZAINAL; MAAROF, 2013; BHOJANNAWAR; BULLA; DANAWADE, 2013; ZHANG; MERATNIA; HAVINGA, 2010):

- **Centralizada**: na detecção centralizada, a estação base realiza o processo de detecção. Após os nodos da RSSF coletarem e enviarem os dados para a estação base, é então que estes dados serão processados e analisados. Normalmente uma estação base possui mais recursos disponíveis para o processamento dos algoritmos de detecção. A capacidade de armazenamento geralmente é maior, possibilitando um maior registro dos dados coletados. Em contrapartida, esta abordagem pode causar uma sobrecarga no tráfego de informações na rede, uma vez que cada nodo precisa enviar suas informações para a estação base. Outro problema é o consumo de energia, sendo que a comunicação requer mais energia que o processamento de dados.

- **Distribuída**: na abordagem distribuída, o processo de detecção utiliza um modelo de referência local de *outliers*. O modelo é enviado para um *cluster head* ou estação base, com o objetivo de definir um modelo de referência global, partindo dos modelos de referências locais, que é difundido entre todos os nodos da RSSF. A partir do modelo de referência global é feita a detecção de *outliers*.

- **Local**: nesta abordagem, o processo de detecção de *outliers* ocorre no próprio nodo individualmente. Para isso, deve existir uma colaboração entre os nodos da rede através da correlação espaço-temporal. Ou seja, os nodos mantêm históricos de suas leituras e/ou também utilizam os valores sensoriados pelos nodos vizinhos para identificar as anomalias. É um modelo escalável permitindo a expansão da RSSF para larga escala.

### 2.2.1 Classificação das Técnicas de Detecção de *Outlier* para RSSFs

Segundo Zhang, Meratnia e Havinga (2010), as técnicas de detecção de *outliers* em RSSFs podem ser divididas segundo suas características em seus domínios de aplicação, os tipos de dados processado, como ilustrado na Figura 4.

Estas técnicas estão divididas em: baseadas em estatísticas, baseadas no vizinho mais próximo, baseadas em clusterização (agrupamento), baseadas em classificação e baseadas em decomposição espectral. As técnicas baseadas em estatísticas ainda se dividem em paramétricas e não paramétricas. As paramétricas dividem-se em gaussianas e não gaussianas, enquanto as não paramétricas dividem-se em baseadas no estimador densidade *kernel* e histograma.

A abordagem baseada em classificação se divide em: máquina de vetores de suporte (SVM) e redes bayesianas. As redes bayesianas subdividem-se em: redes *naive bayesian*, *bayesian belief* e *dynamic bayesian.* Por último, a técnica baseada em decomposição espectral, divide-se em componente principal (ZHANG; MERATNIA; HAVINGA, 2010).

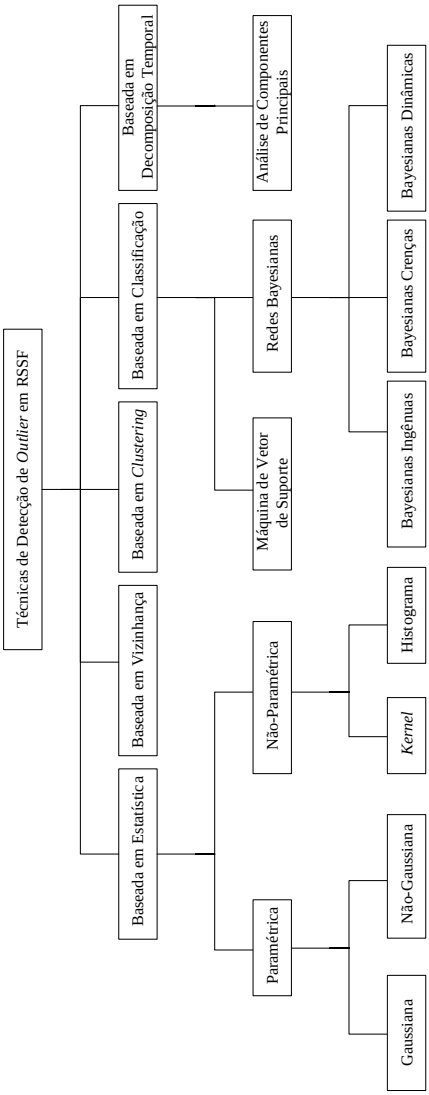
#### 2.2.1.1 Técnicas de detecção de *outliers* baseadas em estatística

As primeiras abordagens desenvolvidas para detectar *outliers* foram com técnicas baseadas em estatísticas. Este tipo de abordagem utiliza modelos para construir um modelo estatístico normal (distribuição de probabilidade), o qual representa o modelo de referência dos dados. Qualquer leitura que o valor desvie do modelo de referência é considerado como uma leitura anômala (ZHANG, 2010; RASSAM; ZAINAL; MAAROF, 2013).

As técnicas baseadas em estatísticas são divididas, de acordo com a disponibilidade do conhecimento prévio do modelo de referência, em paramétricas e não paramétricas.

- **Abordagem Paramétrica**: Nesta abordagem, assume-se que a distribuição dos dados é conhecida *a priori* e os parâmetros da distribuição são facilmente estimados. Com base no tipo de distribuição podemos ainda classificar essas técnicas em gaussiana e não-gaussiana. O modelo gaussiano assume uma distribuição normal dos dados, enquanto que a não-gaussiana são utilizadas

Figura 4 – Técnicas de detecção de *outliers* em RSSFs.




Fonte: Zhang, Meratnia e Havinga (2010).

outras funções de densidade para estimação (ZHANG, 2010; RASSAM; ZAINAL; MAAROF, 2013).

- **Abordagem Não-Paramétrica**: Nesta abordagem, a distribuição dos dados não é conhecida *a priori*. Técnicas que realizam estimativas são utilizadas, como histogramas e estimador de densidade *kernel* para criar um modelo normal de referências que caracterize o comportamento normal dos dados. Técnicas baseadas em histogramas fazem a contagem da frequência de ocorrência dos valores lidos pelos sensores, estimando assim a probabilidade de ocorrência dessa leitura, comparando com cada categoria no histograma e avaliando se pertence a alguma delas (RASSAM; ZAINAL; MAAROF, 2013). As técnicas baseadas no estimador de densidade *kernel* utilizam funções *kernel* para estimar a probabilidade da função distribuída para leituras normais.

As técnicas de detecção de *outliers* baseadas em estatísticas são matematicamente eficientes quando o modelo normal de distribuição de probabilidade é construído corretamente. Entretanto, em aplicações de tempo real nem sempre é conhecido o modelo *a priori*, devido à dinâmica das RSSFs. Com isso, as técnicas não-paramétricas são úteis devido as suas características de não precisarem de conhecimento *a priori*.

O modelo histograma é eficiente apenas para dados univariados, pois não é capaz de capturar interações entre diferentes atributos (multivariados). Enquanto que a função *kernel* pode ser escalável para dados multivariados e possui baixo custo computacional.

#### 2.2.1.2 Técnicas de detecção de *outliers* baseadas em vizinhança

As abordagens baseadas em vizinhança são amplamente utilizadas na área de mineração de dados e aprendizado de máquinas. Essas abordagens utilizam distâncias de similaridade entre os valores sensoriados para medir o grau do padrão normal dos dados ou o grau de anomalia. As funções mais utilizadas são a distância Euclidiana (para valores univariados) ou a distância de Mahalanobis (para valores multivariados). Uma instância de dado é declarada como anômala quando esta está com valores muito distantes dos seus vizinhos (SHAHID; NAQVI; QAISAR, 2015).

Estas abordagens possuem desvantagens em relação ao consumo computacional devido ao processamento em relação aos vizinhos. Por-

tanto, não possui uma boa escalabilidade, ou seja, não é adequada para redes de larga escala. Outra desvantagem é o *overhead* de comunicação (ZHANG, 2010).

#### 2.2.1.3 Técnicas de detecção de *outliers* baseadas em *clustering*

As abordagens baseadas em *clustering* são amplamente estudadas em mineração de dados. Neste tipo de abordagem, os dados são agrupados em *clusters* com base em suas distâncias de similaridade e comportamentos similares. Uma instância é considerada anômala se ela não pertence a um *cluster* ou se um *cluster* é significantemente menor que os outros (SHAHID; NAQVI; QAISAR, 2015).

Cada nodo constrói um modelo de referência local e envia para o *cluster head* (nodo líder do *cluster*). Este constrói um modelo de referência global a partir dos modelos de referência locais. Os *clusters heads* enviam para todos os membros do *cluster* o modelo de referência global acordado. Dessa forma, cada nodo pode detectar anomalias localmente (BHOJANNAWAR; BULLA; DANAWADE, 2013).

Esta técnica não precisa do conhecimento da distribuição dos dados *a priori*. Segundo Rassam, Zainal e Maarof (2013), técnicas de detecção de *outliers* baseadas em *clustering* possuem algumas desvantagens:

- A dependência da escolha da largura dos *clusters* em algumas técnicas de *clustering* inviabiliza o uso em RSSFs.

- *Clustering* é computacionalmente caro especialmente com entrada de dados multivariados, devido ao cálculo das distâncias de similaridade entre os dados. O alto custo computacional é uma limitação intrínseca das RSSFs.

- Possuem limitações para tratar com as mudanças contínuas de fluxos de dados ao longo do tempo, implicando em um modelo de referência normal desatualizado.

#### 2.2.1.4 Técnicas de detecção de *outliers* baseadas em classificação

Técnicas baseadas em classificação se constituem em uma abordagem importante na comunidade de aprendizado de máquina e mineração de dados. Um classificador é treinado, utilizando padrões de

dados de treinamento conhecidos para classificar padrões desconhecidos em um ou mais tipos (RASSAM; ZAINAL; MAAROF, 2013).

Os classificadores aprendem os padrões normais a partir de um conjunto de instâncias de dados e classificam os dados de acordo com o padrão. Se os dados não se encaixam dentro dos limites estabelecidos, estes são considerados como *outliers*(RASSAM; ZAINAL; MAAROF, 2013; ZHANG, 2010).

Um classificador não supervisionado para múltiplos modelos de classes não é ideal para RSSFs pela dificuldade em obter o modelo para diferenciar os dados em normais e anômalos. Entretanto, um classificador não supervisionado com somente um modelo de classe é indicado para RSSFs, devido à utilização de apenas um padrão normal(RASSAM; ZAINAL; MAAROF, 2013).

As técnicas baseadas em classificação são dividas em máquina de vetor de suporte e redes bayesianas:

- **Máquina de Vetor de Suporte**: *Support Vector Machine* (SVM) ou Máquina de Vetor de Suporte. Nessa técnica, os vetores de dados são mapeados para um espaço de característica de alta dimensão (hiperplano) usando funções *kernel*. O modelo que caracteriza os dados normais é encontrado no espaço de alta-dimensão, e os dados anômalos são classificados como sendo aqueles pontos que se desviam do modelo normal neste espaço (RAJASEGARAR; LECKIE; PALANISWAMI, 2008).

- **Redes Bayesianas**: utilizam modelos gráficos probabilísticos para representar um conjunto de variáveis e suas independências probabilísticas. Informações de diferentes variáveis são agregadas para gerar uma estimativa se determinada instância de dados pertence ou não a classe(ZHANG, 2010). As redes Bayesianas ainda se dividem em três categorias de grau de independência entre as variáveis: bayesianas ingênuas, bayesianas baseadas em crenças e bayesianas dinâmicas.

As técnicas baseadas em classificação não supervisionadas não requerem conhecimento dos dados de treinamento, rotulados disponíveis, e aprendem o modelo de classificação que se encaixa na maioria das instâncias de dados durante o treinamento. As técnicas não supervisionadas de uma classe ainda aprendem o limite em torno das instâncias normais. O classificador pode precisar atualizar-se para acomodar a nova instância que pertence à classe normal.

#### 2.2.1.5 Técnicas de detecção de *outliers* baseadas em decomposição temporal

Técnicas de detecção de *outliers* baseada em decomposição temporal buscam encontrar padrões de comportamento nos dados, utilizando análise de componentes principais (PCA). PCA é uma técnica utilizada para reduzir a dimensionalidade dos dados antes da detecção de *outliers*. O objetivo é encontrar um novo subconjunto de dimensão que represente o comportamento dos dados. Essa técnica é computacionalmente de alto custo (ZHANG; MERATNIA; HAVINGA, 2010).

### 2.2.2 Descrição das Técnicas Baseadas em Estatísticas para Detecção de *Outliers* para RSSFs

Abaixo são discutidas técnicas para detecção de *outliers* baseadas em estatísticas. Estas técnicas destacam-se por serem de baixo custo computacional e já serem utilizadas em outras áreas do conhecimento. Além de outras características como adaptabilidade aos cenários dinâmicos, aplicáveis a RSSFs de larga escala e também essas técnicas não precisarem de informações prévias do cenário.

#### 2.2.2.1 Método de Chauvenet

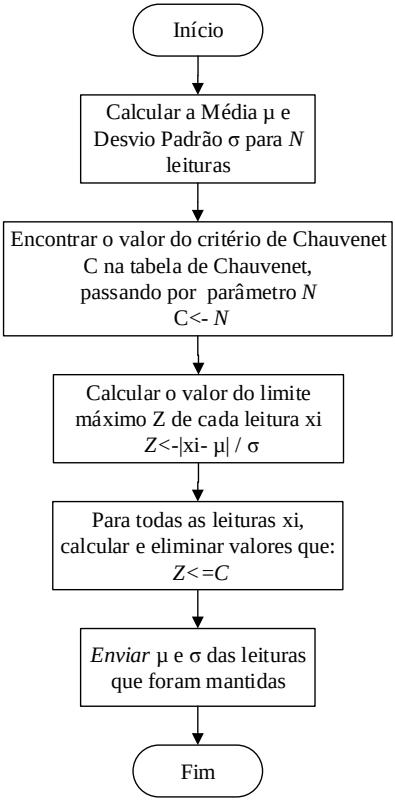
Segundo Taylor (2012), o critério de Chauvenet é utilizado para verificar se em $N$ medições de uma mesma grandeza física podem existir medidas suspeitas de serem diferentes das demais observações. O critério funciona da seguinte forma: deve-se calcular a média e o desvio padrão para todas as $N$ leituras, como consequência pode-se calcular a diferença das amostras suspeitas. O próximo passo é determinar a probabilidade dos valores que estarão fora da faixa definido por 1-1/(2n), ou seja, encontrar os valores considerados como *outliers*.

A Figura 5 apresenta o fluxograma do funcionamento do algoritmo do método de Chauvenet. Inicialmente calcula-se a média $\mu$ e o desvio padrão $\sigma$ para $N$ leituras. A próxima etapa é encontrar o valor do critério de Chauvenet através da Tabela 2, informando a quantidade de leituras do conjunto $N$ como parâmetro. Para cada leitura do conjunto $N$ deve-se calcular o valor do limite máximo $Z$ dado por $|\text{xi-}\mu\ |/\sigma$. Com o valor de Z encontrado, é possível comparar para cada leitura xi, se o seu valor está acima do critério do Chauvenet. Caso esteja a lei-

tura é removida do conjunto. Após a comparação de todas as leituras é calculada a nova média e desvio padrão das leituras mantidas.

A fim de minimizar o processo computacional do cálculo do critério de Chauvenet por integração matemática da função de densidade de probabilidade da distribuição normal, utiliza-se a tabela de *look-up* (Tabela 2) para acessar através do números de amostras, o valor referente ao coeficiente ou critério de Chauvenet (SOARES, 2013).

Figura 5 – Fluxograma método de Chauvenet.




A escolha do tamanho da amostra é muito importante, pois se esta for muito grande a média dificilmente será afetada de forma significativa por uma leitura discordante. Nesse caso, o valor precisará ser extremamente divergente para alterar a distribuição normal.

Tabela 2 – Critério Chauvenet.

| **Tamanho da amostra N** | **Máximo "C" (em desvios)** |
|---|---|
| 3 | 1,38 |
| 4 | 1,54 |
| 5 | 1,65 |
| 6 | 1,73 |
| 7 | 1,80 |
| 8 | 1,87 |
| 9 | 1,91 |
| 10 | 1,96 |
| 15 | 2,13 |
| 20 | 2,24 |
| 25 | 2,33 |
| 50 | 2,57 |
| 100 | 2,81 |
| 300 | 3,14 |
| 500 | 3,29 |
| 1000 | 3,48 |

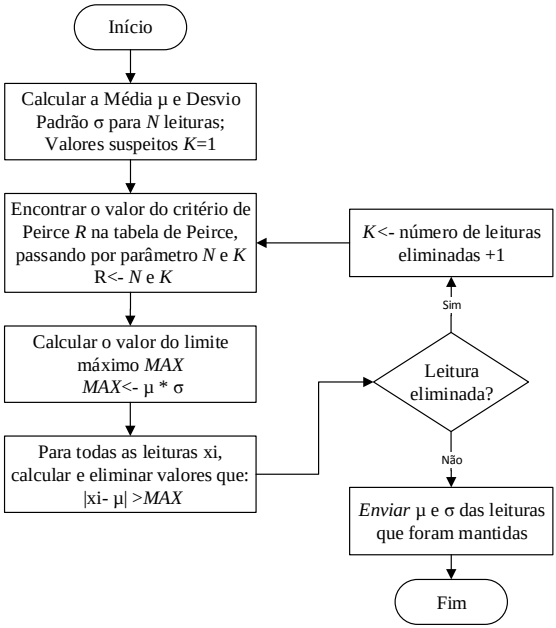
Fonte: Soares (2013).

#### 2.2.2.2 Método de Peirce

Segundo Ross (2003), o método de detecção de *outliers* baseado em probabilidades proposto por Peirce em 1852 afirma que: "as observações devem ser rejeitadas quando os desvios da média obtidos forem menores do que os desvios obtidos por sua rejeição, multiplicada pela probabilidade de fazer tantas, e não mais, observações anormais".

Na Figura 6 é apresentado o fluxograma do funcionamento do método de Peirce. Para um conjunto de $N$ leituras deve-ser calcular a média $\mu$ e o desvio padrão $\sigma$. Em seguida é encontrado o valor do critério de Peirce utilizando a Tabela 3 com os seguintes parâmetros: quantidade de leituras $N$ e quantidade de valores suspeitos $K$. O valor de $K$ inicialmente é 1, mas a cada iteração, e se encontrada uma leitura suspeita, o $K$ é incrementado em 1. Em seguida, é calculado o limite máximo $MAX$ através do cálculo $MAX = \mu^*\sigma$, dessa forma pode-se verificar para toda leitura xi-$\mu$ se o seu valor está abaixo. Se o valor da leitura for maior que MAX, ela é removida do conjunto. Ao final é calculada a média e também o desvio padrão das leituras restantes.

Figura 6 – Fluxograma método de Peirce.




O algoritmo de Peirce possui um custo computacional elevado para calcular o critério. Logo também é utilizada uma tabela de *look-up* com

Tabela 3 – Critério de Peirce.

| **Quantidade de leituras** | **Número de *outliers*** | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| 3 | 1.196 | | | | |
| 4 | 1.383 | 1.078 | | | |
| 5 | 1.509 | 1.200 | | | |
| 6 | 1.610 | 1.299 | 1.099 | | |
| 7 | 1.693 | 1.382 | 1.187 | 1.022 | |
| 8 | 1.763 | 1.453 | 1.261 | 1.109 | |
| 9 | 1.824 | 1.515 | 1.324 | 1.178 | 1.045 |
| 10 | 1.878 | 1.570 | 1.380 | 1.237 | 1.114 |

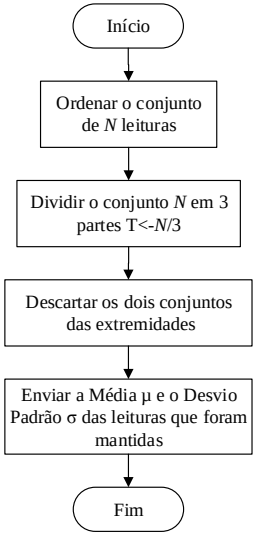
Fonte: Ross (2003).

base no tamanho do conjunto de leituras e da quantidade de possíveis *outliers*. A Tabela 3 contém os valores do critério de Peirce para as primeiras 10 amostras.

#### 2.2.2.3 *Fault Tolerant Averaging*

O método proposto por Marzullo (1990) conhecido como *Fault Tolerant Averaging* (FTA) consiste em ordenar o conjunto de dados (N) e dividir este conjunto em três partes (t=N/3) e a partir disso excluir os conjuntos extremos. Como resultados, calcula-se o desvio padrão e média das leituras restantes como apresenta o fluxograma da Figura 7.

Figura 7 – Fluxograma método FTA.



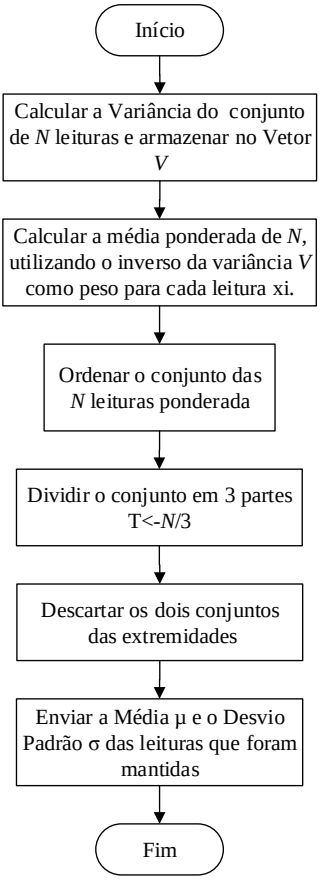

#### 2.2.2.4 *Confidence Weighted Averaging + Fault Tolerant Averaging*

O método *Confidence Weighted Averaging* (CWA) proposto por Elmenreich (2007) é baseado na confiança e variância. Quanto menor a variância das leituras, maior a confiança do sensor. Deve-se aten-

tar para um possível problema quando a variância for muito próxima de zero, pois a mesma pode significar um mau funcionamento do sensor (sensor "travado" em um determinado valor). Assim como uma variância muito alta pode indicar mudanças no comportamento normal esperado (ELMENREICH, 2007).

O CWA sozinho não detecta *outliers*. Para este fim deve associar o CWA a outros métodos de detecção. Neste trabalho, a técnica CWA foi associada ao método FTA. Na Figura 8 é apresentado o fluxograma do CWA+FTA.

Figura 8 – Fluxograma método CWA+FTA.



Primeiro, deve-se calcular, para todo o conjunto de $N$ leituras, a variância e armazenar no vetor $V$. Calcula-se também a média ponderada utilizando o inverso da variância $V$ como peso associado a cada leitura xi. O resultado deverá ser ordenado pela média ponderada. Logo após o conjunto é dividido em três partes e excluem os dois conjuntos da extremidade. Então, a média e o desvio padrão são calculados com as leituras restantes.

### 2.2.3 Classificação das Técnicas de Identificação de *Outliers* para RSSFs

Na literatura, dados com erros e eventos são valores que estão significativamente diferentes, quando comparados com o padrão normal dos dados ou com o restante do conjunto de dados. A principal diferença entre eles é que o primeiro ocorre com maior frequência e normalmente é local, causado por problemas no nodo, enquanto que eventos são globais, ou seja, os nodos vizinhos também irão apresentar leituras semelhantes (BAHREPOUR et al., 2009).

Segundo Kalayci et al. (2011), Rassam, Zainal e Maarof (2013), a detecção de eventos é uma das funções principais das RSSFs. Eventos são as mudanças das grandezas físicas no ambiente monitorado, tal como, derramamento químico, incêndios, mudanças climáticas entre outras.

De acordo com Bahrepour et al. (2009), Pei et al. (2014), Yin, Hu e Yang (2009), a detecção de eventos pode ser classificada em três categorias:

- **Baseada em limites**: A ocorrência de um evento é caracterizada quando as leituras excedem limites predefinidos. Os limites são definidos pelas características da aplicação.
- **Baseada em padrão**: O evento é detectado quando padrões encontrados através da correlação espaço-temporal são identificados. Os padrões precisam ser determinados previamente para cada aplicação.
- **Baseada em aprendizagem de máquina**: Identifica eventos por meio da dependência espaço-temporal dos dados e por inferência probabilística de ocorrer o evento.

Na bibliografia, a identificação de dados com erros e de eventos é tratada por diferentes técnicas. É de grande importância realizar a

distinção entre eles. Basicamente, enquanto os dados com erros são estocasticamente não relacionados, os eventos possuem relação espaço temporal (FAWZY; MOKHTAR; HEGAZY, 2013).

Zhang, Meratnia e Havinga (2010) destacam as principais diferenças entre detecção de eventos e detecção de dados com erros são:

- As técnicas de detecção de dados com erros não mantêm conhecimento *a priori*, enquanto que técnicas de detecção de eventos mantêm um conhecimento prévio sobre a condição de disparo do evento.
- Detecção de dados com erros tem como objetivo identificar dados anômalos comparando as leituras dos sensores entre si, enquanto que detecção de eventos visa analisar as leituras dos sensores com um padrão pré-definido ou uma situação específica.

**Contribuição da detecção e identificação de *outliers* em outras áreas**

Corroborando para a importância das pesquisas com o objetivo de detectar e identificar *outliers*, abaixo seguem algumas aplicações que vem sendo utilizadas, (CHANDOLA; BANERJEE; KUMAR, 2009; ZHANG; MERATNIA; HAVINGA, 2010).

- **Detecção de fraude**: detectar atividades criminosas em organizações comerciais como, bancos, companhias de cartão de crédito, agências de seguros, telefonia entre outros. A fraude ocorre quando usuários autorizados, ou não, consomem os recursos disponibilizados pelas organizações de forma indevida. A identificação imediata de tais fraudes é importante para evitar perdas econômicas.
- **Monitoramento de pacientes**: Os pacientes possuem diversos sensores pelo corpo que monitoram os sinais vitais. A detecção de algum *outlier* pode indicar o surgimento de uma doença ou mudança no estado clínico do paciente. Este tipo de aplicação requer um alto grau de precisão.
- **Detecção de danos industrial**: os equipamentos e estruturas de uso contínuos nas indústrias sofrem desgaste ao longo do tempo. Detectar esses desgastes antecipadamente evitam danos materiais e até mesmo ambientais.
- **Monitoramento de habitat**: realiza o monitoramento de espécies ameaçadas de extinção. Um sensor é colocado no

animal a fim de monitorar o seu comportamento bem como o de seu grupo.

## 2.3 CONSIDERAÇÕES DO CAPÍTULO

Este capítulo envolveu conceitos fundamentais sobre RSSFs e *outliers*. Apresentou a primeira RSSF denominada SOSUS, o padrão IEEE 802.15.4, desenvolvido especialmente para dispositivos de baixo custo e limitações de *hardware*. Salientou as aplicações das RSSFs e a IoT, bem como questões técnicas das estrutura dos nodos e sensores.

Definições fundamentais sobre *outliers* e a classificação das técnicas de detecção e identificação de *outlier* foram discutidas na segunda parte do capítulo. As características e as fontes de origem dos *outlier* também foram estudados. Por fim, foram apresentadas algumas aplicações utilizadas no mundo real.

Com o levantamento do referencial bibliográfico foi possível verificar que a utilização das RSSFs é uma tendência que vem se popularizando, resultado disso foi a padronização do protocolo IEEE 802.15.4.

A respeito das técnicas de detecção de *outliers*, podemos afirmar que as técnicas baseadas em vizinhanças geralmente não são aplicáveis para cenários de larga escala devido ao alto custo computacional, além do *overhead* de comunicação entre os vizinhos. As técnicas baseadas em *clustering* não necessariamente precisam do conhecimento previamente da distribuição dos dados, entretanto, possuem desvantagens como: a determinação do tamanho dos *clusters* e o alto custo computacional para processar dados multivariados.

As técnicas baseadas em classificação apresentam modelos de classificação para detecção de *outliers* precisos. Entretanto, as SVM possuem alta complexidade computacional, além da dificuldade de se manter o modelo de classificação atualizado. As técnicas baseadas em decomposição espectral são aplicáveis em RSSFs de larga escala, entretanto, definir os principais componentes e a matriz de correlação é computacionalmente caro.

Sobre as técnicas de detecção de *outliers* apresentadas, é possível concluir que as baseadas em estatísticas são as que melhores se adaptam às RSSFs, por serem escaláveis, de baixo custo computacional e com a opção de não necessitar de conhecimento *a priori* dos dados.

Nas classificações das técnicas de identificação de *outliers*, destaca-se a baseada em aprendizagem de máquina, por não precisar necessariamente do conhecimento prévio dos dados, enquanto que as

outras duas abordagem necessitam. Entretanto, a baseada em limites e padrão possuem vantagens em relação do baixo custo computacional e poderem ser executadas localmente.

# 3 TRABALHOS RELACIONADOS

Neste capítulo são apresentados os trabalhos correlatos sobre detecção e identificação de *outliers* em RSSFs de larga escala. Como visto na Seção 2.2 existem diversas abordagens para a detecção e identificação de *outliers* em RSSFs. Com o objetivo de padronizar a escolha dos trabalhos analisados, optou-se por usar o método de Revisão Sistemática da literatura (RSL), selecionando os trabalhos mais relevantes da literatura. O objetivo dessa RSL é estabelecer o estado da arte e, assim, abordar os principais problemas e soluções sobre o tema deste trabalho.

A RSL abrangeu seis bases de buscas indexadas com trabalhos publicados entre os anos de 2000 e 2017. Foram utilizadas as seguintes palavras-chave e sinônimos: *outlier*, *anomaly*, *event*, *detection*, *classification*, *identification* e *wireless sensor network*. A descrição completa da execução do protocolo da RSL está incluída no Apêndice A. No total são apresentados 19 trabalhos relacionados como resultado dessa busca sistemática.

O capítulo está organizado em cinco seções. A primeira seção discute os trabalhos correlacionados que realizam apenas a detecção de *outliers*. Os trabalhos que realizam a identificação dos *outliers* são apresentados na segunda seção. A terceira seção reúne os trabalhos que detectam e identificam *outliers* nas mesma abordagem. A quarta seção apresenta uma análise dos trabalhos relacionados com a abordagem proposta. Por fim, a quinta seção traz as considerações deste capítulo.

## 3.1 TRABALHOS CORRELATOS COM DETECÇÃO DE *OUTLIERS*

Nesta seção são discutidos doze trabalhos relacionados que realizam apenas detecção de *outliers*. Os trabalhos são:

- **Rassam, Maarof e Zainal (2014)**: Nesse trabalho foram propostos dois modelos para detecção de anomalias chamados de PCCAD (*Principal Component Classifier-based Anomaly Detection*) e APCCAD (*Adaptive Principal Component Classifier-based Anomaly Detection*) para ambientes estáticos e dinâmicos respectivamente. As duas abordagens utilizam OCPCC (*One-Class Principal Component Classifier*) para mensurar a dissimilaridade entre as leituras sensoriadas.

O modelo PCCAD possui duas fases principais: uma de treinamento e outra de detecção. A fase de treinamento ocorre *offline*, com dados coletados por um determinado período de tempo. A fase de detecção é *online* a cada nova leitura.

O modelo APCCAD possui três fases, além da fase de treinamento e detecção como a do modelo PCCAD. Ele possui uma fase de atualização, em que o modelo é novamente treinado para gerar um novo modelo de referência. Sendo assim, esta abordagem é adaptativa às mudanças do cenário. A abordagem adaptativa APCCAD utiliza um método de aprendizagem incremental para acompanhar as mudanças nos fluxos de dados do ambiente dinâmico monitorado.

As duas abordagens trabalham com modo de detecção de anomalias *online* e utilizam dados multivariados e se mostraram mais eficientes quando comparadas com outras abordagens, como as de origem centralizada. O modelo PCCAD apresentou vantagens por ser de baixa complexidade computacional, entretanto o comportamento em ambientes dinâmicos não teve um bom aproveitamento. Nesse sentido, foi sugerida abordagem adaptativa APCCAD.

- **Moshtaghi et al. (2014)**: Nessa abordagem é proposto um modelo adaptativo para detecção de *outliers* baseado em *clusters* elípticos que definem limites de decisões entre dados normais e dados anômalos. É utilizada uma fração inicial dos dados sensoriados para mapear a distribuição normal de cada sensor, chamado de período de estabilização. A distância de Mahalanobis[1] foi usada para medir a similaridade entre dois vetores de dados sensoriados a fim de formar os *clusters*. Também é proposto um método robusto para modelar o comportamento normal da rede.

  O modelo proposto é adaptativo, portanto ele é continuamente atualizado com as novas leituras obtidas. Os autores afirmam que este modelo pode ser utilizado por diferentes aplicações e cenários. Um problema dessa abordagem é o alto custo de comunicação.

- **Cheng e Zhu (2015)**: Nesse trabalho foram apresentadas duas abordagens para detecção de *outliers* em RSSF denominadas de LADQS (*Lightweight Anomaly Detection Algorithm Using Quick Select*) e LADS (*Lightweight Anomaly Detection Algorithm Using*

[1] A distância de Mahalanobis é uma medida de dissimilaridade entre dois vetores de dados com a mesma distribuição (THE..., 2000).

*Sort*). Ambas as técnicas se assemelham a QSSVM (*One Class Quarter Sphere Support Vector Machine*), entretanto elas apresentam baixa complexidade computacional.

A maior contribuição deste trabalho foi apresentar um método matemático para transformar o problema de otimização linear do QSSVM para um problema de ordenação (*sort*) para diminuir a complexidade computacional dos algoritmos sem diminuir a acurácia na detecção de *outliers*.

- **Rajasegarar et al. (2009)**: Esse trabalho propõe um método baseado em hiper-elipsoides para modelar o comportamento normal das leituras sensoriadas. Ele utiliza também a distância de Mahalanobis para calcular o modelo local padrão dos dados.

  Neste trabalho foram definidas três categorias de anomalias elípticas, sendo elas: de primeira ordem, segunda ordem e alta ordem. Também foram definidos três tipos de detecções elípticas: *Elliptical Cardinality Anomalies, Elliptical Chi-squared Anomalies, e Elliptical Number of Sigma Anomalies.*

  A técnica utiliza um modelo de estrutura distribuído e, segundo os autores, o algoritmo detecta anomalias com a mesma precisão que uma estrutura centralizada com a vantagem de diminuir o consumo energético da rede.

- **Xie, Hu e Guo (2015)**: Essa abordagem é baseada na manipulação dos dados em segmentos. O algoritmo detecta anomalias por meio da segmentação dos dados coletados pelos vizinhos, explorando as correlações espaciais. Essas correlações são realizadas com um detector de predição de variância.

  Para cada segmento de dados, são feitas observações ao longo de um intervalo de tempo contínuo. Um dado segmento é dito como anômalo se contiver múltiplas leituras fora do modelo cêntrico estatístico modelado.

  A abordagem encontrou um problema no custo de processamentos das matrizes de covariância, entretanto foi encontrada uma maneira de minimizar esse problema com a utilização de um coeficiente de correlação. A proposta, quando comparada com trabalhos centralizados, conseguiu atingir uma redução de até 80% do custo de comunicação da rede.

- **Gil, Santos e Cardoso (2014)**: Esse trabalho propõe uma abordagem para detecção de *outliers* para um cenário industrial

de uma refinaria de óleo com restrições temporais *hard* com monitoramento de valores univariados.

A abordagem proposta utiliza inteligência artificial. Um *framework* de multiagentes com gráficos de controle de Shewhart[2] foi implementado para definir limites superiores e inferiores, observando uma única variável para determinar esses limites. Os dados que ficarem de fora desses limites são rotulados dados anômalos. Os limites são baseados nas médias dos dados brutos.

A arquitetura da hierarquia dos multiagentes é composta de duas camadas: camada superior, destinada às funcionalidades do coordenador e a camada inferior, responsável por especificar as tarefas dos agentes como: monitoramento e conversão das leituras para analógico digital, detecção de anomalias entre outros. Cada agente trabalha de forma local, podendo tomar decisões de quando iniciar, pausar e parar a execução das tarefas.

- **Amidi, Hamm e Meratnia (2013)**: Essa proposta foi elaborada sob um cenário de uma estação meteorológica localizada entre a Itália e a Suíça, chamada de passagem *Grand Saint Bernard*. Foram utilizados dados reais para a validação da proposta.

  A metodologia ocorre em três etapas: a primeira fase é a definição dos padrões utilizando informação contextual do cenário de dados já coletados anteriormente. A segunda fase é a avaliação das similaridades e a fase final é a detecção dos *outliers*. Uma desvantagem dessa abordagem é a utilização de dados já coletados anteriormente, necessitando de um histórico da aplicação.

- **Jayashree, Arumugam e Vijayalakshmi (2007)**: Nesse trabalho é apresentada uma técnica estatística chamada de *z-score*, que foi modificada para detectar *outliers* em RSSFs. Esta técnica utiliza o desvio padrão absoluto da média (MAD) para etiquetar de forma confiável os *outliers* e removê-los.

  O processo de remoção é dado em dois momentos antes da identificação do dado anômalo e após a identificação. Primeiramente, são eliminadas as leituras com os valores mais extremos, tanto superior e inferior. Em seguida, é aplicado o método *z-score* modificado para identificar os *outliers* e, então, estes podem ser removidos do conjunto. Uma vantagem desta abordagem é que ela é aplicável em cenários de larga escala.

[2]Shewhart é uma ferramenta gráfica e analítica para monitorar variações do processo.

- **Zhang et al. (2013)**: Nesse trabalho é proposta a união de três modelos: modelo Gaussiano multivariado, análise de componentes principais PCA (*Principal Component Analysis*) e funções do *kernel*. O primeiro modelo analisa as alterações dos dados, correlacionando essas mudanças com um determinado *cluster* de sensores, utilizando modelo Gaussiano.

  O segundo modelo captura a relação geométrica entre os sensores da RSSF com o modelo PCA, pois nem todos os sensores participam de *clusters*. Entretanto, este modelo possui limitações, pois, se a natureza dos dados possuir características não lineares, o modelo PCA não será capaz de capturar os dados corretamente.

  O último modelo utiliza funções *kernel* para corrigir as limitações do PCA e para mapear os dados para um espaço dimensional para então aplicar o modelo de PCA nos dados linearizados.

- **Bosman et al. (2017)**: Nesse trabalho é proposto um modelo distribuído que realiza a fusão dos dados dos vizinhos, visando minimizar a comunicação. O processo de agregação dos dados acontece nos dados sensoriados dos nodos vizinhos. Este processo é feito por meio da sumarização dos dados através da utilização de um operador de agregação, como média, desvio padrão, mediana ou mínimo e máximo.

  Para realizar a detecção são exploradas as correlações espaço temporais das leituras obtidas dos sensores vizinhos. A questão da quantidade ideal de vizinhos também é abordada.

- **Zhang (2010)**: Essa abordagem foi baseada em *One Class Quarter Sphere Support Vector Machine* respeitando as limitações das RSSFs. Estendendo o modelo de QSSVM, os autores propuseram quatro métodos para detecção de *outliers*. O modelo é adaptativo, ou seja, recebe atualizações para acompanhar as mudanças da distribuição dos dados.

  A primeira abordagem classifica de forma *online*, cada nova leitura em normal ou anômala. As outras três abordagens são para atualizar o modelo de representação do comportamento normal dos dados, levando em consideração as correlações temporais entre as leituras mais recentes. Espera-se que, com a aplicação de QSSVM, ocorra uma redução da complexidade computacional.

- **Rajasegarar et al. (2014)**: Nesse trabalho os autores propuseram uma arquitetura de detecção de anomalias para RSSFs. A

arquitetura utiliza diversos *clusters hyperellipsoidal* para modelar os dados de cada nodo e, assim, identificar as anomalias globais bem como as locais na RSSF.

Além disso, é proposto um método para distribuir uma pontuação para cada nova *hyperellipsoidal* baseado na distância de cada elipse com as elipses vizinhas. Os resultados demonstram diminuição do *overhead* de comunicação na rede quando comparada com técnicas de estrutura centralizada.

## 3.2 TRABALHOS CORRELATOS COM IDENTIFICAÇÃO DE *OUTLIER*

Nesta seção são discutidos dois trabalhos relacionados que realizam apenas identificação de *outliers*. Os trabalhos são:

- **Cao et al. (2014)**: Nesse trabalho, um modelo *online* e distribuído para detecção de eventos em RSSFs foi proposto. Esta técnica não necessita de conhecimento prévio do modelo da rede.

  A abordagem utilizada foi de sensoriamento comprimido, ou seja, utilizou-se o algoritmo de programação linear OMP para detectar as anomalias modeladas como *l1-norm minimization problem*[3]. Este processo é iterativo a fim de capturar o estado atual do ambiente ou fenômeno sensoriado.

  O processo de identificação dos eventos é global. O algoritmo OMP atribui pesos as leituras. Em seguida, esses pesos são comparados com um limiar e assim é decidido se possui ou não um evento, caso os pesos estejam fora do valor do limiar. Os pesos são atualizados para se adaptar às mudanças da rede.

- **Wang et al. (2015)**: O objetivo do trabalho de Wang et al. (2015) foi propor uma abordagem distribuída para detecção de eventos em RSSFs utilizando o método de *self learning threshold* (limiar de auto-aprendizagem).

  Primeiramente, a sequência de fluxo de dados é transformada em sequências simbólicas, com o objetivo de reduzir a dimensionalidade dos dados e simplificar a identificação dos eventos. Através do modelo de Markov são estimadas as probabilidades de conter anomalias nas sequências de símbolos.

[3] A *l1-norm minimization problem* é um problema de otimização convexa, basicamente minimiza a soma das diferenças entre um valor dado e o valor esperado.

Cada nodo pode contribuir para a criação do limiar, dessa forma pode-se aprender/estimar um novo limiar a cada sequência de leituras. Os autores ainda propõem um esquema de escalonamento de sono para aumentar a vida útil da RSSF.

## 3.3 TRABALHOS CORRELATOS COM DETECÇÃO E IDENTIFICAÇÃO DE *OUTLIER*

Nesta seção são discutidos cinco trabalhos que realizam detecção e identificação de *outliers* na mesma abordagem ou arquitetura. Os trabalhos são:

- **Wang, Lin e Jiang (2014)**: Nesta proposta, os autores apresentaram um método de detecção de *outliers* e eventos usando o modelo oculto de Markov (*Hidden Markov*) de multi dimensões. Esse modelo é aplicado em RSSF para capturar a correlação entre as dimensões de dados.

  A proposta é baseada na detecção da trajetória dos *outliers*, ou seja, detecção dos comportamentos dos *outliers* utilizando um modelo de treinamento e também um modelo baseado em estimativa de probabilidade. Para as mudanças de comportamento, foi proposta também uma alteração para os cenários dinâmicos. As correlações e detecção são realizadas em tempo real.

  A abordagem proposta possui baixo custo computacional para realizar o processamento de dados, clusterização treinamento e atualização do modelo. Os resultados mostraram que o algoritmo foi capaz de identificar mudanças de comportamento, distinguindo entre dados anômalos e eventos.

- **Wu et al. (2007)**: Este trabalho teve como objetivo a proposição de dois algoritmos para detecção de *outlier* juntamente com a detecção de eventos em RSSFs localmente. Os autores propuseram dois algoritmos escalável.

  As técnicas utilizam correlação espacial das leituras entre os vizinhos para identificar se é o *outlier* um dado com erro ou um evento. Para detectar *outlier*, cada nodo deve calcular a diferença entre suas próprias leituras com a média das leituras vizinhas. O nodo é considerado com dados anômalos se o valor da diferença das médias for suficientemente maior que um limiar pré-definido.

Para identificar eventos, é analisado se o grau de diferença dos nodos de uma região geográfica é muito maior que de outra região. Os autores apontam que, como não é utilizada correlação temporal, a precisão da técnica não é muito alta.

- **Fawzy, Mokhtar e Hegazy (2013)**: A proposta apresentada para detectar *outliers* e eventos é baseada na abordagem de *clustering* combinada com abordagem baseada do vizinho mais próximo.

  A proposta é composta por quatro etapas: a primeira fase é o pré-processamento dos dados sensoriados, ou seja, os dados são agrupados em *clusters*. A segunda fase é a da detecção de *outliers*, onde, para cada *cluster*, é aplicado o algoritmo de detecção de *outliers* para rotular quais dados são anômalos ou não.

  Na terceira fase é feita a identificação do tipo do *outlier*, se é um evento ou um erro/ruído, através das correlações dos vizinhos mais próximos. A última etapa atribui um peso de confiabilidade para cada sensor, de forma que, quando um nodo apresentar um *outlier* identificado como ruído, este terá um peso mais baixo.

- **Shahid, Naqvi e Bin Qaisar (2014)**: Para identificar *outliers* e eventos, os autores propuseram a utilização de SVM (*Support Vector Machines*) em RSSFs estendendo a abordagem para QS-SVM (*Quarter Sphere Support Vector Machine*) para detectar dados anômalos. Na identificação de eventos, os autores utilizam o *SensGru*[4], ou seja, um nodo que possui vários sensores embutidos a ele mesmo. Desta maneira, evita-se a comunicação desnecessária entre os nodos sem comprometer as vantagens da correlação espacial temporal e de atributos.

- **Salem et al. (2013)**: Esse trabalho propõe um *framework* para detecção de anomalias em RSSF na área médica. Apresenta-se um cenário onde vários sensores estão conectados ao paciente e deseja-se monitorar os sinais vitais do paciente e transmitir os dados coletados em um intervalo regular de tempo para um *smartphone* (estação base).

  Essa abordagem utiliza duas técnicas para realizar as correlações: distância de Mahalanobis para realizar a análise espacial e KDE

[4]SensGru é nomenclatura dada aos nodos que possuem múltiplos sensores pelos autores

(*kernel density estimator*) para análise temporal. Além da correlação espaço-temporal também é levado em consideração os atributos fisiológicos dos pacientes.

O principal objetivo dessa aplicação é diferenciar uma leitura anômala de uma emergência médica em tempo real. Este trabalho mostrou bons resultados na detecção de anomalias e uma baixa complexidade computacional. Entretanto, como foi utilizada uma abordagem centralizada para a comunicação dos dados, o consumo energético da rede foi alto.

## 3.4 ANÁLISE DOS TRABALHOS RELACIONADOS COM ABORDAGEM PROPOSTA

Nesta seção é feita uma comparação entre a abordagem proposta e as abordagens encontradas na literatura. Como um dos resultados da RSL temos a Tabela 4 com os trabalhos mais relevantes e pertinentes ao tema deste trabalho que abordam detecção e/ou identificação de *outliers* em RSSF de larga escala.

Os atributos de comparação das propostas levam em consideração as características de cada abordagem proposta por cada autor. Essas características foram escolhidas com base nos trabalhos de Rassam, Zainal e Maarof (2013) e Zhang, Meratnia e Havinga (2010), são elas: detecção e identificação de *outliers*, se o modelo de estrutura é local, centralizado ou distribuído, se a dimensão dos dados é univariada e multivariada, se o modo de operação é *online* ou *offline*, se há análise de correlação temporal, espacial ou de atributos, se a abordagem é escalar ou não.

Ainda na Tabela 4, temos as técnicas de detecção utilizadas por cada abordagem, classificadas pela taxonomia apresentada na Figura 4; Além das técnicas de identificação classificadas segundo a Seção 2.2.3.

A proposta realizada neste trabalho destaca-se das outras na questão de realizar a detecção e identificação de *outliers* na mesma abordagem, enquanto que as maiorias das propostas abordam essa temática de maneira independente. Além disso, outras características são consideradas, como o modelo de estrutura local para minimizar os custos com comunicação, a utilização da correlação espacial e temporal. Quando se trata de uma RSSF com grandes quantidades de nodos, a abordagem precisa ser escalável, enquanto maioria das propostas não considera essa questão. Além disso, a proposta utiliza técnicas para detecção e identificação de *outliers* de baixo custo computacional.

Tabela 4 – Comparação da proposta com os trabalhos relacionados da RSL.

| Abordagem | Ruídos e Erros | Eventos | Modelo de Estrutura | | | Dimensão dos dados | | Modo de Operação | | Correlação | | | Escalar | Técnica Utilizada Para Detecção de Ruídos e Erros | Técnica Utilizada Para Identificação de Eventos |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Local | Centralizada | Distribuída | Univariado | Multivariado | Offline | Online | Espacial | Temporal | Atributos | | | |
| Cao et al. (2014) | | ✓ | | | ✓ | | ✓ | | ✓ | | | | | n/a | Baseada em Limites |
| Moshtaghi et al. (2014) | ✓ | | | | ✓ | | ✓ | ✓ | | | | | | Baseada em Clusterização | n/a |
| Wang et al. (2015) | | ✓ | | | ✓ | | ✓ | | ✓ | ✓ | | | | n/a | Baseada em Aprendizado de Máquina |
| Cheng e Zhu (2015) | ✓ | | | | | | ✓ | | | | | | | Baseada em Classificação | n/a |
| Rajasegarar et al. (2009) | ✓ | | | | ✓ | | ✓ | ✓ | | | ✓ | | ✓ | Baseada em Clusterização | n/a |
| Xie, Hu e Guo (2015) | ✓ | | | | ✓ | | ✓ | | ✓ | ✓ | | | ✓ | Baseada em Estatística | n/a |
| Salem et al. (2013) | ✓ | ✓ | ✓ | | | | ✓ | | ✓ | ✓ | ✓ | ✓ | | Baseada em Estatística | Baseada em Padrões e Aprendizado de Máquina |
| Wang, Lin e Jiang (2014) | ✓ | ✓ | | | ✓ | | ✓ | | ✓ | ✓ | ✓ | | | Baseada em Estatística | Baseada em Aprendizado de Máquina |
| Gil, Santos e Cardoso (2014) | ✓ | | ✓ | | | ✓ | | | ✓ | | | | | Baseada em Estatística | n/a |
| Amidi, Hamm e Meratnia (2013) | ✓ | | | | ✓ | ✓ | | ✓ | | ✓ | ✓ | | | Baseada em Estatística | n/a |
| Jayashree, Arumugam e Vijayalakshmi (2007) | ✓ | | | | ✓ | ✓ | | | ✓ | | | | ✓ | Baseada em Estatística | n/a |
| Zhang et al. (2013) | ✓ | | | | | | ✓ | | | ✓ | ✓ | | | Baseada em Estatística e Decomposição Espectral | n/a |
| Rassam, Maarof e Zainal (2014) | ✓ | | ✓ | | | | ✓ | | ✓ | | | | | Baseada em Decomposição Espectral | n/a |
| Wu et al. (2007) | ✓ | ✓ | ✓ | | | ✓ | | | ✓ | ✓ | | | ✓ | Baseada em Estatística | Baseada em Aprendizado de Máquina |
| Bosman et al. (2017) | ✓ | | | | ✓ | | ✓ | | ✓ | ✓ | | | ✓ | Baseada em Vizinhança | n/a |
| Fawzy, Mokhtar e Hegazy (2013) | ✓ | ✓ | | | ✓ | | ✓ | | ✓ | ✓ | ✓ | | | Baseada em Clusterização | Baseada em Aprendizado de Máquina |
| Zhang, Meratnia e Havinga (2010b) | ✓ | | | | ✓ | | ✓ | | ✓ | ✓ | ✓ | | | Baseada em Classificação | n/a |
| Rajasegarar et al. (2014) | ✓ | | | | ✓ | | ✓ | | ✓ | ✓ | | | | Baseada em Clusterização | n/a |
| Shahid, Naqvi e Bin Qaisar (2014) | ✓ | ✓ | | | ✓ | | ✓ | | ✓ | ✓ | ✓ | ✓ | | Baseada em Classificação | Baseada em Aprendizado de Máquina |
| **Proposta** | ✓ | ✓ | ✓ | | | ✓ | | | ✓ | ✓ | ✓ | | ✓ | Baseada em Estatística | Baseada em Limites e Padrões |

Legenda: n/a significa não aplicável.

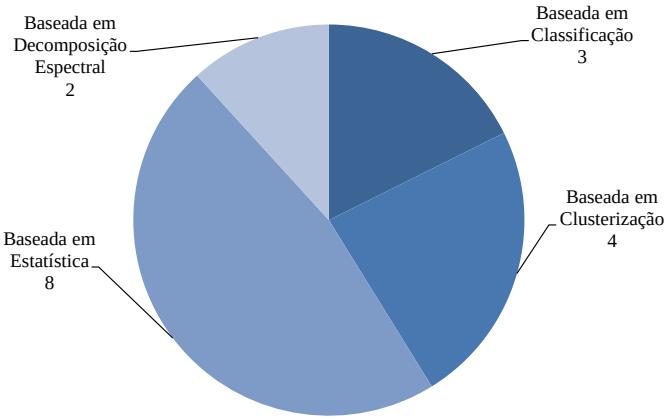
## 3.5 CONSIDERAÇÕES DO CAPÍTULO

Como resultado da RSL obteve-se todos os trabalhos relevantes sobre o tema deste trabalho no estado da arte. A utilização de um protocolo de busca para encontrar os trabalhos relacionados permite documentar todo o método executado. Além de possibilitar que outras pessoas realizem a mesma busca seguindo os mesmos critérios e encontrem os mesmos artigos como resultados.

Foram levantados com a RSL diversos trabalhos sobre detecção e identificação de *outliers* na literatura. Entretanto, poucos trabalhos contemplam os dois processos na mesma abordagem. O foco de estudo da maioria dos trabalhos é propor técnicas para detecção de *outliers*.

A RSL ainda corroborou para apontar que a utilização das técnicas baseadas em estatísticas são as predominantes em RSSFs. Dos 17 trabalhos que realizam detecção de *outliers*, 8 utilizam técnicas baseadas em estatísticas, 4 utilizam técnicas baseadas em clusterização, 3 utilizam técnicas baseadas em classificação e 2 trabalhos utilizam técnicas baseadas em decomposição espectral, conforme a Figura 9.

Figura 9 – Trabalhos relacionados por classificação das técnicas de detecção de *outliers*




A utilização das técnicas baseadas em estatística justifica-se pelo baixo custo computacional, o que é favorável para os cenários de RSSFs

e as restrições de recursos de *hardware* dos nodos. Além disso, normalmente essas técnicas também não precisam de conhecimento prévio da rede se adequando ao cenário dinâmico das RSSFs.

# 4 ABORDAGEM PARA DETECÇÃO E IDENTIFICAÇÃO DE *OUTLIERS* EM RSSFs DE LARGA ESCALA

Neste capítulo é apresentada a abordagem para detecção e identificação de *outliers* em RSSF de larga escala. O capítulo está divido em quatro seções: a primeira seção apresenta uma visão geral da proposta, a segunda seção é dedicada a descrição da proposta e está dividida em mais duas subseções. Na terceira seção são descritos os critérios de avaliação para a proposta, por fim temos a quarta seção com as considerações do capítulo.

## 4.1 VISÃO GERAL DA PROPOSTA

Como já discutido anteriormente no Capítulo 2, é importante manter os dados brutos gerados por RSSFs confiáveis para que eles possam ser utilizados em futuras tomadas de decisões. Através da detecção e identificação de *outliers* podemos garantir que o sistema estará livre de dados anômalos. Caso dados anômalos sejam detectados, a identificação deles é muito importante, pois um evento pode estar por trás da sua geração.

Na literatura estão disponíveis diversos trabalhos (como os discutidos na Seção 3.1) que tratam sobre a detecção de *outliers* separadamente da sua identificação. A maioria das abordagens elimina os *outliers* dos conjuntos de dados. Entretanto, essas abordagens nem sempre são as mais apropriadas, pois os *outliers* podem ter sidos gerados por um evento (ex. incêndios, derramamento químico).

A partir desses problemas, propomos uma abordagem dividida em duas etapas, uma para detecção de *outliers* e outra, com base nos resultados da primeira, para identificação (classificação) dos *outliers* em dados espúrios (dados com erros) ou eventos relevantes. O tratamento dos dados fica a critério da necessidade da aplicação, podendo exclui-los ou mantê-los armazenados em uma base de dados externa.

O objetivo dessa proposta é realizar a detecção e identificação de *outliers* em RSSF de larga escala, monitorando uma única grandeza física. A detecção dos *outliers* ocorrerá de forma local e *online* obedecendo uma restrição temporal *soft*, enquanto a identificação será realizada através de correlações espaço-temporais no coordenador.

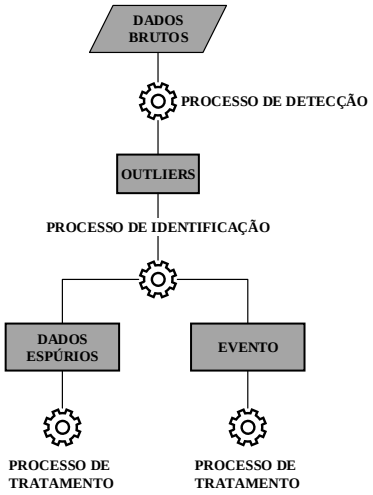
## 4.2 DESCRIÇÃO DA PROPOSTA

Após a implantação da RSSF e sua ativação, a rede começa a coletar dados. Esses dados são chamados de dados brutos, ou seja, não foram analisados ou tratados por nenhuma técnica de tratamento de dados (mineração de dados, processamento de dados entre outras).

À medida que os sensores vão coletando os dados por meio de leituras do ambiente monitorado, é aplicado o processo de detecção, com o objetivo de verificar se o conjunto de dados possui ou não dados anômalos. Caso existam dados anômalos, estes são avaliados pelo processo de identificação para verificar se é um dado espúrio ou evento. Após a identificação do dado anômalo, um processo adequado de tratamento dos dados é essencial.

Na Figura 10 é possível visualizar a interação dos processos de detecção, identificação e tratamento dos dados. Neste trabalho são abordados apenas os processos de detecção e identificação de *outliers*. O tratamento dos *outliers* fica sob a responsabilidade da aplicação da RSSF.

Figura 10 – Etapa de detecção, identificação e tratamento de *outliers* em RSSFs.




Fonte: Adaptado de Andrade (2016).

Em resumo, o processo de detecção é responsável por diferenciar se o dado bruto possui ou não anomalias. Na sequência, a etapa de

identificação recebe o dado anômalo detectado e verifica se este dado é um evento relevante ou dado espúrio.

Abaixo, nas duas próximas seções, é apresentada e discutida a proposta que, para melhor entendimento, está dividida em duas etapas: a primeira trata da detecção dos dados anômalos (seção 4.2.1) e a segunda na identificação, se os dados anômalos são dados espúrios ou resultante de um evento relevante (Seção 4.2.2).

### 4.2.1 Descrição da Abordagem para Detecção de *Outliers*

As técnicas tradicionais para detecção de *outliers* não são adequadas para RSSFs devido ao alto custo computacional requerido para execução dos algoritmos. As técnicas para RSSFs precisam ser de baixo custo computacional respeitando assim as características e limitações dos nodos. As técnicas que mais se encaixam nessas características são as baseadas em estatísticas (ZHANG; MERATNIA; HAVINGA, 2010; RASSAM; ZAINAL; MAAROF, 2013). Neste trabalho são avaliadas quatro técnicas baseadas em estatísticas para detecção de *outliers*. As técnicas selecionadas foram: CWA+FTA, FTA, Chauvenet e Peirce (ver Seção 2.2.2).

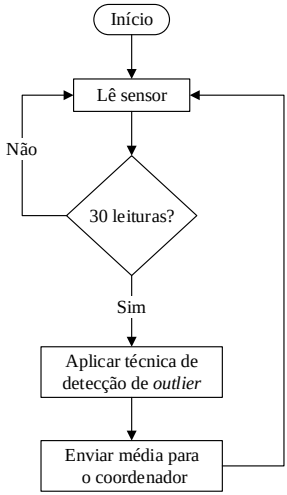
Cada nodo possui o algoritmo de detecção de *outliers* implementado localmente. Como o processo de detecção é local, não é necessário enviar todas as leituras para o coordenador, diminuindo, por consequência, o fluxo de comunicação de dados dos nodos e da rede, bem como o gasto energético. O consumo energético da RSSF não é mensurado neste trabalho, entretanto, essa questão não é negligenciada, pois buscou-se utilizar técnicas com pouco processamento local e que reduzissem a comunicação entre nodos. O quê, de forma indireta, reduz o consumo energético da RSSF.

O modo de operação do processamento é realizado de forma *online*, ou seja, a detecção é realizada imediatamente após a leitura dos dados. Entretanto, a restrição temporal é *soft*, ou seja, se algumas leituras não forem realizadas a tempo, a aplicação não será prejudicada significantemente. A escolha da restrição temporal depende de cada aplicação.

As quatro técnicas de detecção supracitadas são avaliadas, nesta seção, com o objetivo de selecionar a melhor técnica que será implementada em cada nodo da RSSF para realizar o processo de detecção de *outliers*. Os resultados da comparação das técnicas podem ser analisados na Seção 5.3.5.

Na Figura 11 é apresentado o fluxograma do funcionamento do nodo. Cada nodo é implementado com a técnica de detecção que obteve os melhores resultados. O nodo realizará o sensoriamento do ambiente e armazenará 30 leituras (esse valor é arbitrário mas foi escolhido para se ter uma amostra de dados significativa para se calcular localmente os valores de média e desvio padrão). na sequência é aplicada uma técnica de detecção de *outliers* com a finalidade de melhorar a precisão das médias das leituras do nodo, através da eliminação dos dados anômalos (dados espúrios) no próprio nodo. Como resultado da execução da técnica de detecção obtemos uma média das leituras, esta média deverá ser enviada ao coordenador para dar continuidade a segunda etapa da proposta.

Figura 11 – Fluxograma do funcionamento do nodo.




Segundo o fluxo de processos que executam na RSSF apresentado na Figura 10, após o processo de detecção dos *outliers* é necessário identificar se os dados anômalos detectados são dados espúrios ou evento. Para tal, foi proposta a segunda etapa da abordagem que combina duas técnicas de identificação de *outlier* para RSSFs de larga escala (Seção 4.2.2).

### 4.2.2 Descrição da Abordagem para Identificação de *Outliers*

Conforme visto na Seção 2.2, a origem das fontes de geração de *outliers* em RSSFs pode ser, segundo o Zhang, Meratnia e Havinga (2010), classificada em três tipos: eventos, erros ou ruídos e ataques maliciosos. Nesta segunda etapa da abordagem é verificado se o *outlier* detectado na primeira etapa da abordagem é um evento; caso contrário, ele é considerado como dado espúrio.

Para o processo de identificação de *outliers* é proposta a combinação de duas técnicas: uma baseada na vizinhança e outra em limites pré-definidos. Quando dados anômalos são detectados como *outliers* pelo processo de detecção, o coordenador da RSSF verificará se todos os nodos vizinhos aos nodos com *outliers* também apresentaram anomalias em suas leituras, buscando por correlações espaciais. Além disso, o coordenador verificará se as leituras dos nodos vizinhos estão acima do limite pré-definido, caso estejam é identificado um evento. A união das duas técnicas para identificar *outliers* são úteis para evitar falsos alarmes.

O processo de correlação dos nodos é feito com base na localização dos mesmos. O coordenador é responsável por manter atualizada uma tabela de vizinhança para todos os nodos da RSSF. O coordenador gera essa tabela de vizinhança através do indicador da potência do sinal recebido RSSI (*Received Signal Strength Indicator*) de cada nodo. Contudo se a RSSF for equipada como nodos com GPS, dados de vizinhança mais precisos podem ser usado.

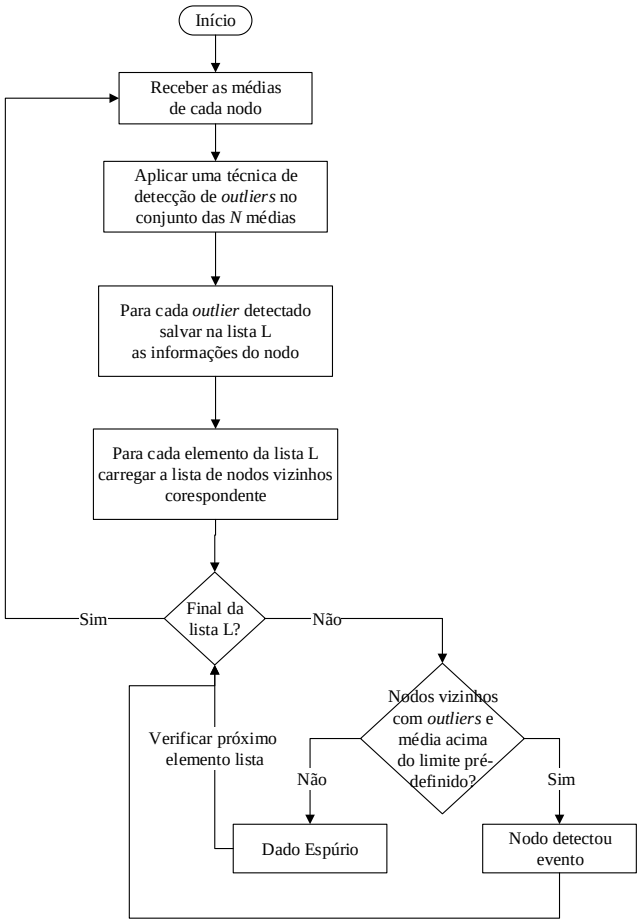
A criação da tabela de vizinhos é muito importante, pois ela é utilizada para realizar a correlação espacial entre os nodos que detectaram dados anômalos e relacionar se os nodos com as mesmas anomalias detectadas também são vizinhos. Caso sejam, um evento relevante pode ser a causa dos *outliers* e atenção a esta área de monitoramento deve ser aumentada. Além da correlação espacial, também é usada a correlação espaço-temporal das leituras, pois mantém-se o histórico das leituras por um determinado período de tempo no processo de detecção de *outliers*.

Todo o processo de identificação ocorre centralizado no coordenador. O coordenador, após um tempo definido, receberá dos nodos uma média das leituras gerada pelo processo de detecção. Com base nas médias, o coordenador verifica com uma técnica de detecção de *outliers* se todas as médias recebidas dos nodos estão dentro do mesmo comportamento. Caso uma ou mais médias não estejam dentro do comportamento esperado, o coordenador verifica se os vizinhos a estes nodos

também possuem anomalias e se possuem valores de médias de leituras acima do limite pré-definido. Se apenas nodos sem correlação estiverem com dados anômalos, estes são classificados como dados espúrios. Caso contrário, um evento relevante foi identificado.

Na Figura 12 é apresentado o fluxograma do funcionamento do algoritmo no coordenador.

Figura 12 – Fluxograma do funcionamento do coordenador.




O fluxograma inicia com o coordenador recebendo as médias das

leituras de todos os nodos da RSSF, para então aplicar uma técnica de detecção de *outliers*. Esta técnica deve discriminar exatamente qual o elemento que foi removido do conjunto das médias recebidas, para que se possa manter o controle dos nodos que possuem *outliers* e posteriormente poder analisar se existe alguma correlação entre os nodos a fim de identificar um possível evento. Como resultado da técnica obtemos uma lista com todos os nodos que apresentaram *outliers*. Para cada elemento dessa lista deve-se verificar os vizinhos correspondentes ao nodo. Caso um nodo da lista possua um ou mais vizinhos que também foram detectados com *outliers* e, além disso, se suas médias de leituras foram maiores que o limite pré-definido, um evento foi identificado, caso contrário, é um dado espúrio.

## 4.3 CRITÉRIOS DE AVALIAÇÃO DAS TÉCNICAS DE DETECÇÃO DE *OUTLIERS*

Devido às características distintas entre as quatro técnicas de detecção de *outliers* selecionadas para a análise, onde as técnicas FTA e CWA+FTA não apontam o elemento excluído e nem a quantidade de *outliers* removidos, enquanto, que Peirce e Chauvenet discriminam o elemento detectado, foi desenvolvido um método para avaliação das técnicas que abrange todas as quatros técnicas, com o objetivo de avaliar de forma quantitativa a técnica que mais detectar *outliers* corretamente.

O método consiste em inserir artificialmente *outliers* no conjunto de dados de leituras dos nodos, com a finalidade de manter o controle de quantos *outliers* espera-se que possam ser detectados. Portanto, a técnica que detectar o maior número de *outliers* artificiais é a escolhida. No caso de empate em relação a quantidade de *outliers* detectados, será utilizada a tabela 5 para desempate.

A tabela 5 foi criada com base das características das técnicas, como a quantidade máxima de leituras que podem ser analisadas a cada iteração e se realiza ou não remoção indiscriminadas dos *outliers*.

Quando houver empate com as técnicas Chauvenet e Peirce, Peirce será escolhida devido esta possuir uma maior quantidade de leituras que podem ser analisadas simultaneamente. Quando Chauvenet obter empate com FTA ou CWA+FTA, FTA ou CWA+FTA será a escolhida. O mesmo ocorre para empate com Peirce e FTA ou CWA+FTA, sendo FTA ou CWA+FTA será a escolhida por não terem restrições nas quantidades de leituras que podem ser analisadas.

Por fim, se ocorrer empate entre FTA e CWA+FTA, CWA+FTA, será escolhida por não realizar remoção indiscriminadas nas leituras.

Tabela 5 – Escolha dos métodos para desempate.

| **Empate entre os métodos** | **Escolha o método** |
|---|---|
| Chauvenet e Peirce | Peirce |
| Chauvenet e FTA | FTA |
| Chauvenet e CWA+FTA | CWA+FTA |
| Peirce e FTA | FTA |
| Peirce e CWA+FTA | CWA+FTA |
| FTA e CWA+FTA | CWA+FTA |

Para avaliação da abordagem de identificação de *outliers* também foi elaborado um método de avaliação. Ele consiste em inserir leituras artificialmente no conjunto de dados de determinados nodos para assim simular a ocorrência de um evento. Por consequência, a abordagem proposta deve ser capaz de realizar a identificação desses dados corretamente como eventos ou dados espúrios quando for o caso. A aplicação desses métodos e critérios para avaliação são apresentados no próximo capitulo.

## 4.4 CONSIDERAÇÕES DO CAPÍTULO

Neste capítulo, a abordagem proposta neste trabalho para detectar e identificar *outliers* em RSSF de larga escala foi apresentada. A abordagem proposta busca se encaixar nas lacunas encontradas com a RSL. Este trabalho oferece uma abordagem de detecção e identificação de *outliers* em dados univariados, com detecção local e *online*, e identificação desses *outliers*, usando correlação espaço-temporal em grande áreas.

O objetivo da abordagem proposta é encontrar *outliers* e identificá-los com sucesso utilizando técnicas de baixo custo computacional. Espera-se que a proposta possa identificar os eventos corretamente sem falsos alarmes. Na etapa de detecção dos *outliers* os dados anômalos já são removidos no próprio nodo. Caso o coordenador receba valores das média dos nodos que não possuam correlação com outros nodos, é entendido que o nodo possui alguma inconsistência ou são dados espúrios. Entretanto, se o nodo possuir vizinhos com as mesmas anomalias, isto indica que um possível evento relevante esteja causando as anomalias.

O próximo capítulo apresenta a avaliação da abordagem proposta em um cenário de larga escala através de simulações.

# 5 AVALIAÇÃO DA ABORDAGEM PARA DETECÇÃO E IDENTIFICAÇÃO DE *OUTLIERS* EM RSSFs DE LARGA ESCALA

O objetivo deste capítulo é avaliar a abordagem apresentada no Capítulo 4, através do uso de um simulador para RSSFs. É apresentado um cenário para os testes de detecção de *outliers* e simulação de eventos. A avaliação da abordagem proposta é dividida em duas etapas. No primeiro momento são avaliadas as técnicas baseadas em estatísticas para detecção de *outliers* e, posteriormente, na segunda etapa avalia-se a proposta para identificação dos *outliers*.

Os resultados obtidos mostraram que é possível utilizar técnicas de baixa complexidade computacional para detectar e identificar *outliers* em RSSFs de larga escala.

O capítulo está divido em cinco seções. A primeira discute o simulador utilizado neste trabalho. A segunda seção descreve o cenário utilizado para a validação da proposta. A terceira seção apresenta os resultados obtidos nas avaliações das técnicas de detecção de *outliers*. A quarta seção apresenta os resultados obtidos na avaliação da abordagem para identificação de *outliers*. A última seção é dedicada as considerações e resultados do capítulo.

## 5.1 SIMULADOR PARA RSSFs: CASTALIA

Para a simulação da RSSF de larga escala foi escolhido o *framework Castalia* 3.0[1]. *Castalia* é um simulador de código aberto desenvolvido com a plataforma OMNeT++[2] para RSSFs e *Body Area Networks*(BANs).

A justificativa para a escolha desta ferramenta é motivada pela grande aceitação e uso na comunidade de RSSFs. Devido às características realísticas para simulação do comportamento do meio de transmissão, modelos de rádio, modelo de bateria e simulação do processo físico para simulação de sensores (BOULIS, 2010). Outra característica desta ferramenta é a possibilidade dos pesquisadores desenvolver seus próprios algoritmos e protocolos. A descrição da RSSF utilizada nesta ferramenta é descrita na seção a seguir.

[1] https://github.com/boulis/Castalia
[2] https://omnetpp.org/

## 5.2 DESCRIÇÃO DO CENÁRIO DAS SIMULAÇÕES

Para a avaliação das abordagens foi utilizada uma RSSF de larga escala. Para tal, foi buscado na literatura um *dataset* que atendia aos requisitos do trabalho: ser de larga escala e possuir armazenamentos de dados de no mínimo uma grandeza física. O *dataset* utilizado para a simulação e avaliação da abordagem foi o gerado pelo projeto LUCE[3] (*Lausanne Urban Canopy Experiment*) mantido pelo projeto Sensorscope[4].

A RSSF do projeto foi implantada no campus da École Polytechnique Fedérale de Lausanne (EFPL), e inicialmente possuía 110 nodos distribuídos por uma área de 300x400 metros, caracterizando segundo Ingelrest et al. (2010) uma RSSF de larga escala. O *download* da base de dados do projeto LUCE pode ser encontrado no *link* "http://lcav.epfl.ch/page-86035-en.html".

O trabalho realizado pelo Sensorscope tinha por objetivo monitorar as mudanças meteorológicas do campus. Foram coletados dados como temperatura, umidade relativa do ar, velocidade e direção do vento, radiação solar, entre outras grandezas físicas. Os dados coletados através da RSSF LUCE correspondem ao período de julho de 2006 a maio de 2007.

Verificando a consistência da base de dados LUCE identificou-se que ocorreu um mau funcionamento de alguns nodos no experimento *in loco*. Portanto, neste estudo foram utilizados um total de 85 nodos (84 nodos para monitoramento e 1 de coordenador/*sink*). A implantação dos nodos pelo campus EFPL pode ser visualizada na Figura 13.

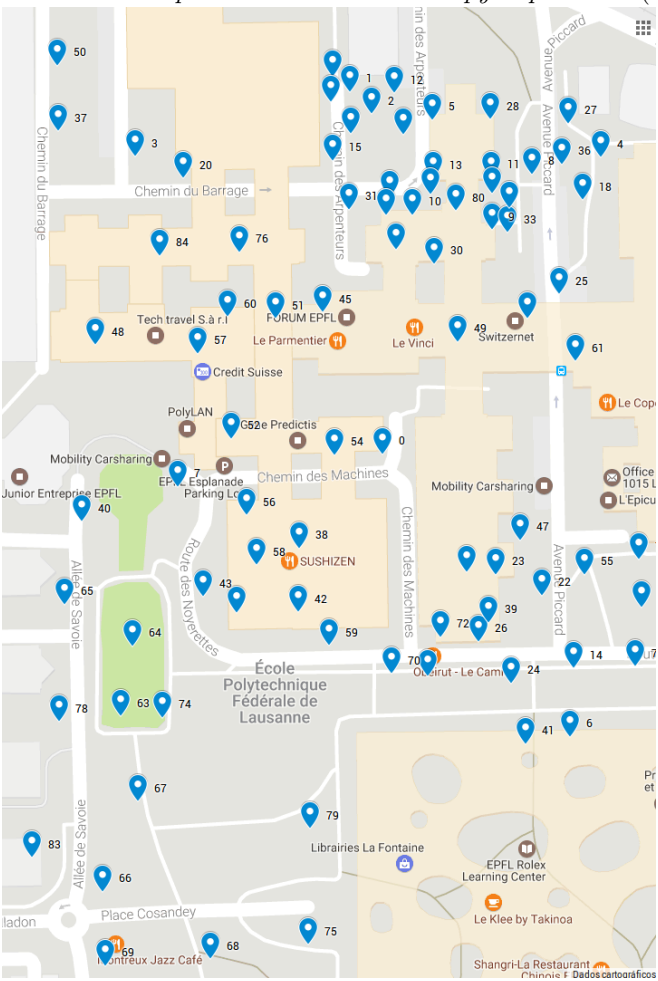
Para limitar a abrangência do trabalho, foi escolhido simular um total de 24 horas de coleta de dados, e também foi utilizada apenas uma grandeza física: temperatura. Os dados são referentes ao dia 03 de abril de 2007 com leituras realizadas a cada 30 segundos. A escolha do valor da periodicidade das leituras foi mantida conforme a utilizada no projeto LUCE.

No simulador Castalia, a RSSF LUCE foi implementada em topologia estrela (um coordenador). Todos o nodos seguem o padrão de rede IEEE 802.15.4 sem *beacon*. O rádio utilizado foi o CC2420, o qual é utilizado em diversos sensores comerciais, como é o caso do MICAz[5]. Como foram utilizados dados reais e já existentes, neste trabalho não houve a pretensão de se aprofundar no estudo sobre o protocolo de co-

[3] http://lcav.epfl.ch/page-86035-en.html
[4] http://lcav.epfl.ch/cms/site/lcav/lang/en/sensorscope-en
[5] http://www.memsic.com/wireless-sensor-networks/

municação dos dados e sincronismo dos nodos. Assume-se que todos os nodos possuem conectividade com o coordenador.

Figura 13 – *Sensorscope Lausanne Urban Canopy Experiment* (LUCE).




Os nodos possuem posições conhecidas e estacionárias, e essas posições foram inseridas no simulador. A posição geográfica utilizada para os nodos é a mesma oferecida pelo projeto Sensorscope. O simulador OMNeT++/Castalia foi alterado neste trabalho de mestrado de forma a permitir a importação de dados para serem utilizados como

leituras dos sensores, emulando um processo físico e permitindo a simulação da comunicação na rede com esses dados.

## 5.3 AVALIAÇÕES DAS TÉCNICAS DE DETECÇÃO DE *OUTLIERS* BASEADAS EM ESTATÍSTICAS

Nesta seção são descritas as simulações realizadas no ambiente *Castalia*. Foram realizadas simulações para encontrar a melhor técnica de detecção de *outliers* respeitando as restrições da RSSF apresentada anteriormente como baixo custo, limitação de processamento e memória entre outros. Também foram realizadas simulações para avaliar a proposta de identificação dos *outliers* apresentada no Capítulo 4.

Foram implementadas as quatro técnicas de detecção de *outliers* baseadas em estatísticas selecionadas no Capítulo 4. E posteriormente foi Parametrizado o cenário das simulações para avaliação das técnicas temos:

- O monitoramento foi realizado a cada 30 segundos e, após a obtenção de 30 leituras, o método de detecção de *outliers* é executado;
- Após o processo de detecção local, cada nodo transmite ao coordenador o resultado da média adquirida;
- De forma local e individual, cada nodo utiliza o algoritmo de detecção de *outliers* com exceção do coordenador.

A escolha de trinta leituras se deve ao fato que, segundo Hogg, Tanis e Zimmerman (2003), a quantidade entre vinte e cinco a trinta amostras é suficiente para representar a distribuição normal de um conjunto de dados baseados no teorema central do limite (HOGG; TANIS; ZIMMERMAN, 2003).

O tempo entre as leituras é flexível e pode ser modificado de acordo com a aplicação, entretanto neste trabalho foi preferível manter os intervalos de tempo utilizados pelo Sensorescope.

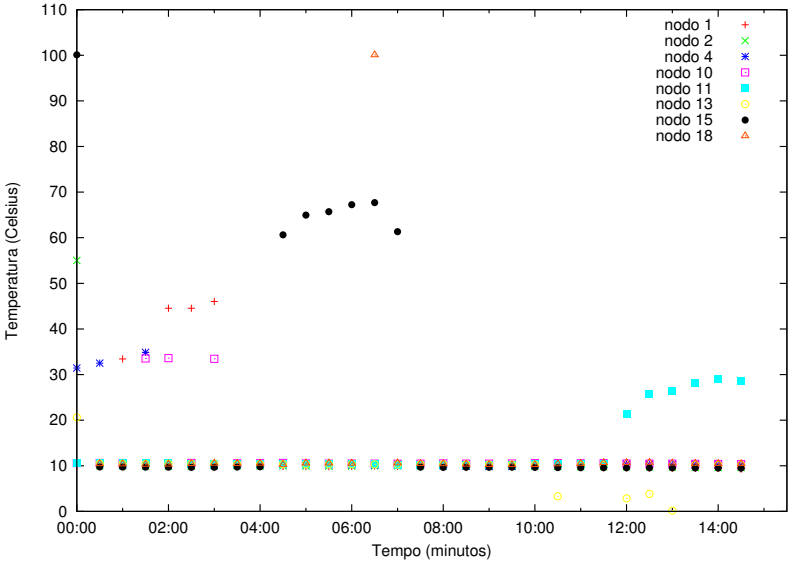
Os métodos avaliados foram: Chauvenet, Peirce, FTA e CWA+FTA. Uma alteração no método CWA+FTA foi feita para possibilitar a detecção de *outlier* no próprio nodo. Ao invés de calcular a confiabilidade dos sensores, por meio da coleta de uma única leitura de todos os sensores, foi realizado o cálculo no conjunto das leituras referentes a um período para cada sensor individualmente. Portanto, como resultado temos uma lista ordenada das leituras por variância, podendo aplicar o método FTA em seguida.

Nas técnicas de Peirce e Chauvenet é possível descobrir a quantidade exata e o valor da leitura dos *outliers* que foram detectados, entretanto, isso não ocorre nas técnicas de FTA e CWA+FTA. No sentido de analisar a eficiência dos métodos de detecção de *outlier*, foram inseridos *outliers* artificiais nos dados do monitoramento (como previsto nas Seção 4.3). A escolha dos valores dos *outliers* artificiais inseridos em cada nodo, foi feita de forma aleatória.

Para analisar os resultados foram selecionados oito nodos (10% do total) com melhores frequências de envios ao coordenador. Os nodos escolhidos foram os que alcançaram 100% de frequência de envios. Ao todo foram distribuídos 30 *outliers* entre os oito nodos.

A Figura 14 apresenta as leituras de temperatura dos nodos no período de quinze minutos, juntamente com os 30 *outliers* artificiais. É possível observar a discrepância dos valores e a necessidade de tratamento para que esses dados anômalos não distorçam o resultado final, alterando o valor real das médias.

Figura 14 – Leituras sem processo de detecção *outliers*.

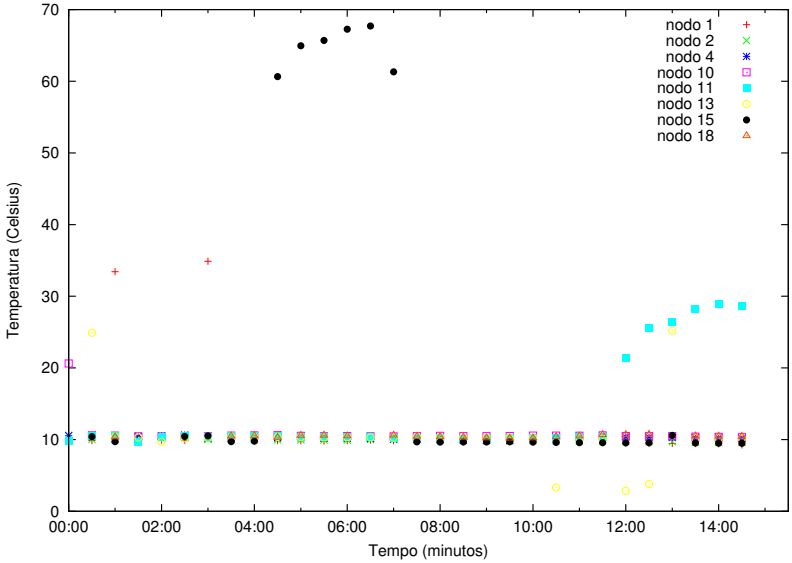



Nas seções abaixo estão descritos os resultados para as quatro técnicas de detecção de *outliers*, analisados individualmente.

### 5.3.1 Chauvenet

O método de detecção de *outliers* proposto por Chauvenet detectou 13 *outliers* dos 30 *outliers* artificiais inseridos, como pode ser visto na Figura 15. Os valores fora da faixa próximo da temperatura de 10ºC, são *outliers*. O gráfico apresenta muitas leituras destoantes do comportamento normal esperado das leituras capturadas. No total foram 17 *outliers* não detectados pela técnica. O aproveitamento desta técnica foi 43,3% de detecção dos *outliers*.

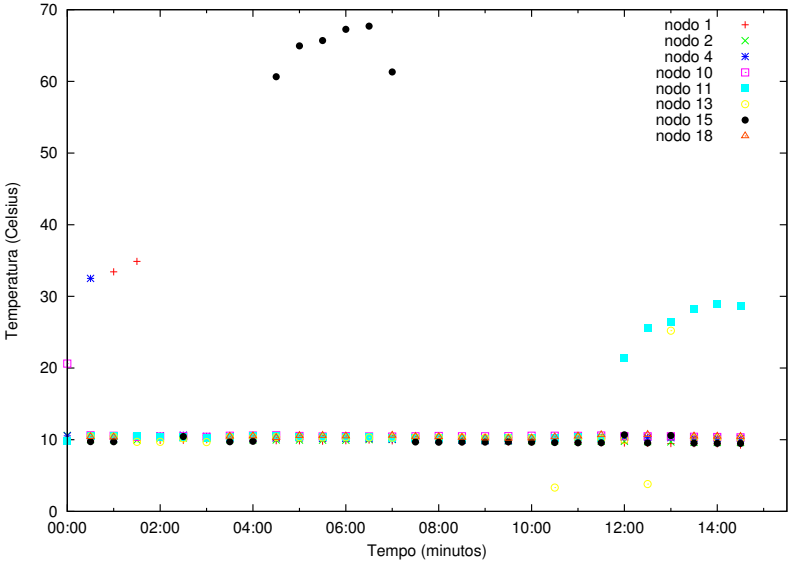
Figura 15 – Resultado da detecção com o método Chauvenet.



### 5.3.2 Peirce

A técnica Peirce apresenta um comportamento muito semelhante ao de Chauvenet para a detecção de *outliers*. Também identificou corretamente apenas 13 *outliers*. Na Figura 16 é possível ver os 17 *outliers* não detectados. O aproveitamento da técnica também foi de 43,3%.

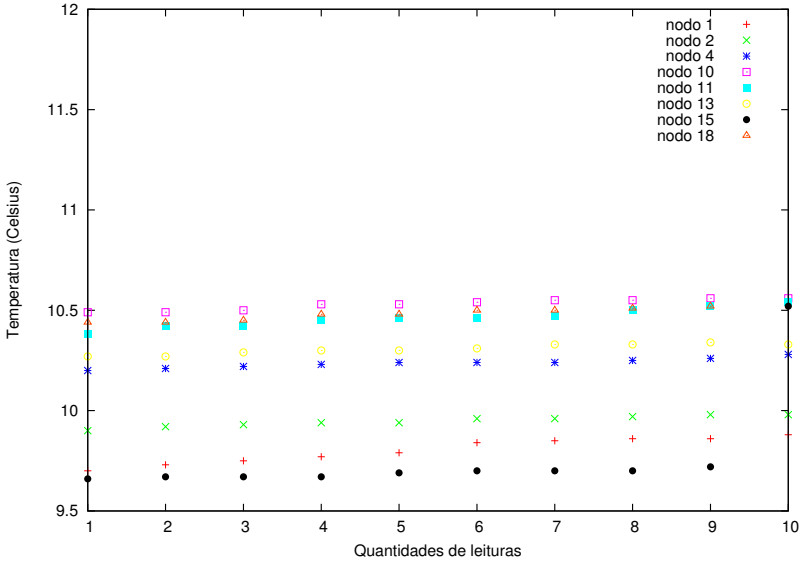
Figura 16 – Resultado da detecção com o método Peirce.




### 5.3.3 FTA

O método FTA mostrou-se muito eficiente na identificação dos *outliers*. Ela foi capaz de encontrar todos os *outliers* inseridos, ou seja, teve 100% de aproveitamento. Apesar de ser um método relativamente simples mostrou ser eficiente para detecção e remoção dos *outliers*. A Figura 17, mostra as leituras resultantes do método FTA. Um problema identificado nesta técnica foi o da remoção indiscriminada dos valores, mesmo aqueles que não são considerados dados anômalos, porque ela sempre excluir 2/3 das amostras.

Figura 17 – Resultado da detecção com o método FTA.



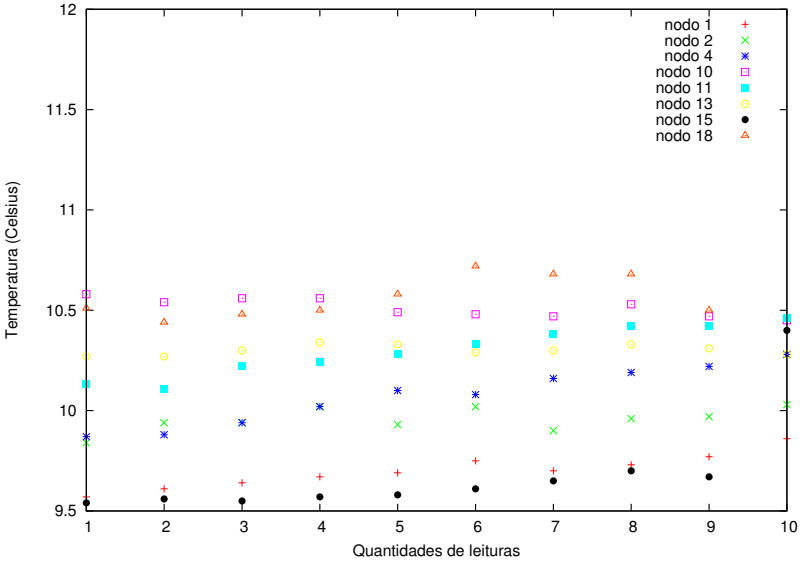

### 5.3.4 CWA+FTA

O método proposto por Elmenreich com a associação do método FTA obteve resultados semelhantes ao método FTA, conforme a Figura 18, alcançando 100% de eficiência na detecção dos *outliers*. O problema do FTA em remover de forma indiscriminada é suavizado nesta técnica, pois o CWA inicialmente realiza uma ordenação pelo peso das variâncias das leituras. Portanto, as leituras duvidosas ficam nas extremidades que serão removidas pelo FTA.

### 5.3.5 Análise Comparativa Entre as Técnicas de Detecção de *Outliers* Baseadas em Estatísticas

Os métodos obtiveram, de modo geral, bons resultados para detecção dos *outliers* no cenário proposto. Contudo, os métodos FTA e CWA+FTA detectaram 100% dos *outliers*, enquanto os métodos Peirce e Chauvenet detectaram apenas 43,3% dos *outliers*. A Tabela 6 apresenta as quantidades de *outliers* detectados por nodo. Na tabela é

Figura 18 – Resultado da detecção com o método CWA+FTA.




possível observar que as técnicas FTA e CWA+FTA obtiveram resultados similares como também as técnicas Peirce e Chauvenet obtiveram resultados parecidos entre as duas.

Tabela 6 – Quantidade de *outliers* detectados por técnica.

| | **Total de *outliers*** | **CWA+FTA** | **FTA** | **Peirce** | **Chauvenet** |
|---|---|---|---|---|---|
| **Nodo 1** | 5 | 5 | 5 | 3 | 3 |
| **Nodo 2** | 1 | 1 | 1 | 1 | 1 |
| **Nodo 4** | 3 | 3 | 3 | 2 | 3 |
| **Nodo 10** | 3 | 3 | 3 | 3 | 3 |
| **Nodo 11** | 6 | 6 | 6 | 0 | 0 |
| **Nodo 13** | 4 | 4 | 4 | 2 | 1 |
| **Nodo 15** | 6 | 6 | 6 | 0 | 0 |
| **Nodo 18** | 2 | 2 | 2 | 2 | 2 |
| **Total** | **30** | **30** | **30** | **13** | **13** |

A Tabela 7, apresenta as médias finais das leituras dos nodos

por método avaliado. Em análise, a média dos valores das leituras dos nodos 11 e 15, após o processo de detecção utilizando as técnicas de Peirce e Chauvenet, apresentaram médias elevadas, devido à presença de *outliers* remanescentes. Essa condição de média elevada, não corresponde à variação de temperatura naquele período para estes nodos. Enquanto que nas técnicas de detecção FTA e CWA+FTA houve uma regularização da média final dos nodos, ou seja, todos *outliers* foram removidos tornando as informações das médias para tomada de decisões mais precisas e confiáveis.

Tabela 7 – Médias por técnica em grau Celsius.

| | **Média** | **CWA+FTA** | **FTA** | **Peirce** | **Chauvenet** |
|---|---|---|---|---|---|
| **Nodo 1** | 14,86 | 9,69 | 9,80 | 11,51 | 11,51 |
| **Nodo 2** | 11,39 | 9,95 | 9,94 | 9,89 | 9,89 |
| **Nodo 4** | 12,46 | 10,07 | 10,23 | 10,98 | 10,18 |
| **Nodo 10** | 12,81 | 10,51 | 10,53 | 10,51 | 10,51 |
| **Nodo 11** | 13,61 | 10,29 | 10,46 | 13,61 | 13,61 |
| **Nodo 13** | 9,24 | 10,30 | 10,30 | 9,81 | 9,56 |
| **Nodo 15** | 20,63 | 9,61 | 9,69 | 20,63 | 20,63 |
| **Nodo 18** | 16,42 | 10,54 | 10,48 | 10,45 | 10,45 |

Em resumo, comparando a capacidade de detecção de *outliers*, tanto a técnica CWA+FTA como o FTA alcançaram os mesmo resultados. Em relação à precisão, as técnicas também obtiveram resultados semelhantes, conforme Tabela 7. Como as técnicas CWA+FTA e FTA foram muito semelhantes nos resultados será utilizado como critério de desempate a complexidade do método. Portanto, a técnica que obteve melhor desempenho neste cenário foi a CWA+FTA.

Com a escolha da técnica CWA+FTA, a segunda etapa da proposta pode ser avaliada, pois esta será a técnica que ficará implementada no nodo localmente para o processo de detecção de *outliers*.

## 5.4 AVALIAÇÃO DA ABORDAGEM PARA IDENTIFICAÇÃO DE *OUTLIERS*

A segunda etapa da avaliação da abordagem proposta consiste na avaliação do processo de identificação de *outlier*. Depois que o processo de detecção com o método CWA+FTA, enviar as médias de leituras de temperatura de cada nodo para o coordenador, o processo de identificação é executado para diferenciar entre dados espúrios e eventos.

O processo de identificação primeiramente detectará com o método de Peirce se as médias recebidas pelo coordenador possuem *outliers*. Após detectar que as leituras possuem *outliers*, necessitamos determinar o tipo desses *outliers*. Se for um evento, poderá ser enviado um alarme para o usuário; caso sejam dados espúrios, esses dados podem ser removidos do conjunto de dados. Importante ressaltar que o tratamento adequado para os dados espúrios é totalmente dependente da aplicação.

A escolha do método de Peirce no coordenador se dá pela necessidade de saber exatamente qual o nodo que possui dados anômalos, para assim poder correlacionar com os nodos vizinhos e verificar se existe ou não a presença de um evento. Outro motivo para a escolha do método de Peirce é da capacidade de calcular um conjunto maior de amostras em comparação do método de Chauvenet, por este motivo ele foi escolhido para ser implementado no coordenador e não o método de Chauvenet. Os métodos FTA e CWA+FTA não poderiam ser utilizados nesta etapa, por não apontarem o elemento exato que foi detectado como *outlier*.

Para a simulação de detecção de eventos, foram inseridos valores que simulavam a ocorrência de um evento. No estudo de caso utilizado neste trabalho focou em um exemplo de incêndio, por motivos de estarmos monitorando a grandeza física de temperatura.

De acordo com a NBR ISO 7240-5:2008, a temperatura estática de resposta é a temperatura na qual a aplicação deverá tomar uma decisão, caso tenha alguma alteração excedendo os limites. Os limites definidos pela NBR ficam entre 54$^{o}$C e 65$^{o}$C (NBR ISO 7240-5, 2008). Foi considerado limite máximo aproximadamente de 60$^{o}$C, devido às limitações dos sensores utilizados pelo Sensorscope. Portanto, as temperaturas lidas não ultrapassam o valor de 60$^{o}$C.

Segundo a empresa especializada em detectores de incêndio ABAFIRE[6], a cada minuto a temperatura sobe oito graus Celsius em curto espaço de tempo durante um incêndio (ABAFIRE, 2017). As leituras inseridas para simular o incêndio seguem essa premissa.

Para avaliação da abordagem foram simulados dois eventos de incêndio. No primeiro, o fogo inicia da direita para a esquerda. Os primeiros nodos a terem suas leituras alteradas foram os nodos 4 e 27 (em vermelho) e depois os nodos em amarelo 36, 18 e 8. No segundo incêndio, a origem foi entre os nodos 26 e 39 assinalados em vermelho atingindo em seguida os nodos em amarelo 72, 77, 70, 24, 14, 22, 23, 82 e 47, conforme ilustrado na Figura 19.

[6] http://abafire.com.br/

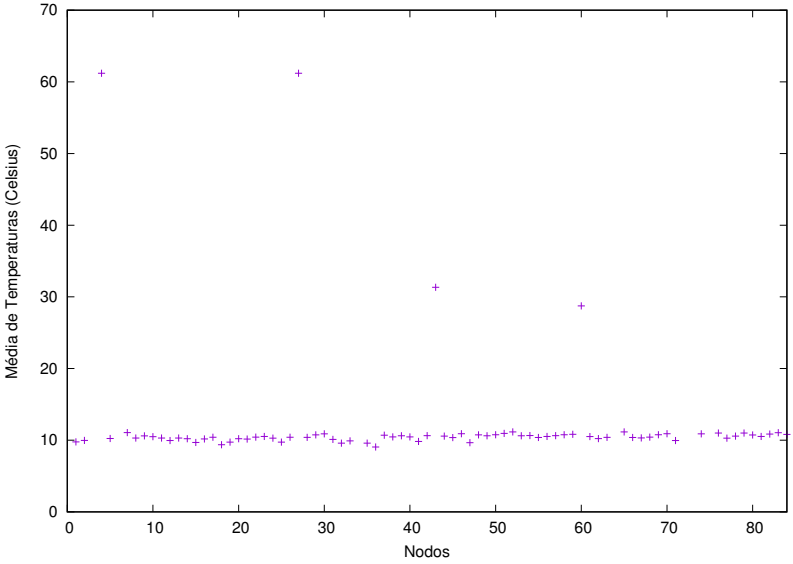
Figura 19 – Cenário de simulação LUCE com eventos inseridos.


Para cada média considerada como *outlier*, o coordenador verifica se os vizinhos deste nodo também apresentaram *outliers*. Caso os vizinhos também apresentem médias identificadas como *outliers* e possuam valores acima do limites pré-definidos, um evento de incêndio

está ocorrendo muito próximo a esses nodos. Entretanto, se o nodo não tiver correlação espacial com nenhum outro nodo vizinho, ele é considerado com dados espúrios.

A Figura 20 apresenta um gráfico com as médias das temperaturas de cada nodo para um determinado instante de tempo. É possível analisar que existem quatro pontos distantes do restante dos dados. Estes pontos são referentes aos nodos 27, 4, 43 e 60.

Figura 20 – Média de temperatura por nodos na primeira iteração.




Para exemplificar melhor, a Tabela 8 ilustra como a abordagem atua. O método de Peirce detectou que os nodos 27 e 4 possuíam médias fora do padrão do conjunto recebido, ou seja, foram detectados como *outliers*. Portanto, é necessário identificar se existe alguma correlação entre esses nodos e as médias da vizinhança.

A Tabela 8 apresenta os nodos 27 e 24 com os valores de suas médias e os valores das médias dos vizinhos. O nodo 27 possui média das leituras em 61,24ºC e possui como vizinho o nodo 4 com média de 61,19ºC, como pode ser conferido na linha Correlações do Nodo 27. Verificando o nodo 4, que foi também detectado com um valor de *outlier*, o mesmo possui correlação com o nodo 27.

Tabela 8 – Identificação de *outliers* primeira rodada.

| | | Nodos Vizinhos | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Nodo** | **27** | 4 | 8 | 11 | 18 | 28 | 35 | 36 |
| Médias ($^{o}$C) | 61,24 | 61,19 | 10,29 | 10,29 | 9,37 | 10,40 | 9,59 | 9,05 |
| **Correlações do Nodo 27** | | X | | | | | | |
| **Nodo** | **4** | 8 | 18 | 27 | 36 | | | |
| Médias ($^{o}$C) | 61,19 | 10,29 | 9,37 | 61,24 | 9,05 | | | |
| **Correlações do Nodo 4** | | | | X | | | | |
| **Nodo** | **43** | 42 | 53 | 56 | 58 | 64 | | |
| Médias ($^{o}$C) | 31,33 | 10,63 | 10,61 | 10,52 | 10,76 | - | | |
| **Correlações do Nodo 43** | | | | | | | | |
| **Nodo** | **60** | 45 | 51 | 57 | 76 | 84 | | |
| Médias ($^{o}$C) | 28,73 | 10,36 | 10,95 | 10,62 | 10,99 | 10,80 | | |
| **Correlações do Nodo 60** | | | | | | | | |

A premissa é que um nodo precisa ter pelo menos um vizinho[7], e se este vizinho apresentar *outlier* e valor da média acima do limite pré-definido, pode-se afirmar que o *outlier* detectado é um evento relevante.

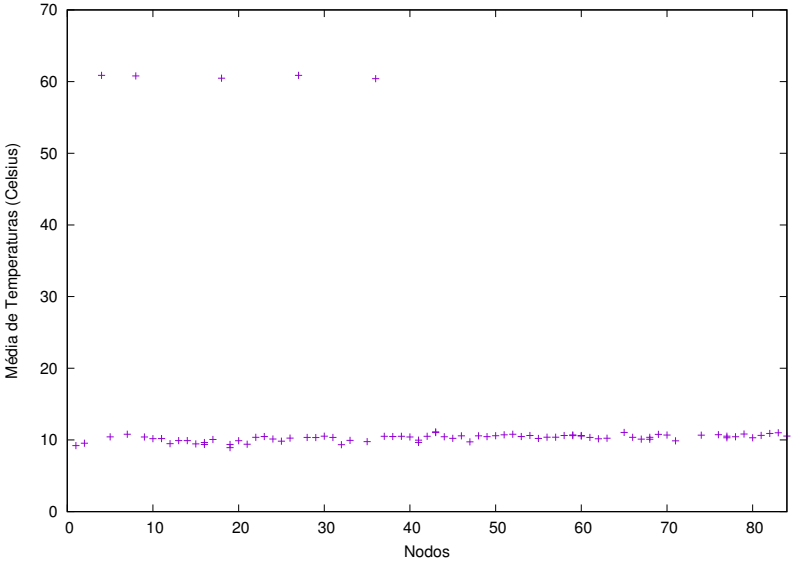
Os nodos 43 e 60 assinalados em preto na Figura 19 foram detectados como *outliers*. Entretanto, essas anomalias não possuem correlações com outros nodos, indicando serem dados espúrios. Como pode ser visto na Tabela 8, os nodos 43 e 60 não apresentam nenhum vizinho

[7]Importante observar que o número mínimo de vizinhos com dados com *outliers* para se detectar um evento é dependente da aplicação, e pode ser configurado no algoritmo.

com as mesmas mudanças nas médias de leituras.

Na próxima iteração, o incêndio continua. O coordenador recebe novas médias de todos os nodos. Como pode ser visto na Figura 19 os nodos 36, 8 e 18 assinalados em amarelo também detectaram valores divergentes. A Figura 21 ilustra o gráfico para a segunda iteração onde existem cinco pontos destoantes.

Figura 21 – Média de temperatura por nodos na segunda iteração.




Na Tabela 9 são apresentadas todas as relações dos nodos que foram detectados como *outliers* pelo método de Peirce. Foram os nodos 4, 36, 8 e 18. Portanto, esses nodos com *outliers* estão detectando alterações de temperatura, resultado de um evento e, assume-se que, não são dados espúrios oriundos de sensores com erros ou ruídos.

Tabela 9 – Identificação de *outliers* segunda rodada.

| | Nodos Vizinhos | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Nodo** | 4 | 8 | 18 | 27 | 36 | | | | | | | | | |
| Médias (ºC) | 60,86 | 60,79 | 60,47 | 60,86 | 60,41 | | | | | | | | | |
| **Correlações do Nodo 4** | | X | X | X | X | | | | | | | | | |
| | | | | | | | | | | | | | | |
| **Nodo** | 36 | 4 | 8 | 9 | 11 | 18 | 27 | 28 | 33 | 34 | 35 | | | |
| Médias (ºC) | 60,41 | 60,86 | 60,79 | 10,41 | 10,18 | 60,47 | 60,86 | 10,35 | 9,93 | - | 9,77 | | | |
| **Correlações do Nodo 36** | | X | X | | | X | X | | | | | | | |
| | | | | | | | | | | | | | | |
| **Nodo** | 8 | 4 | 9 | 11 | 18 | 27 | 28 | 33 | 34 | 35 | 36 | | | |
| Médias (ºC) | | 60,86 | 10,41 | 10,18 | 60,47 | 60,86 | 10,35 | 9,93 | - | 9,77 | 60,41 | | | |
| **Correlações do Nodo 8** | | X | | | X | X | | | | | X | | | |
| | | | | | | | | | | | | | | |
| **Nodo** | 18 | 4 | 8 | 9 | 11 | 27 | 33 | 34 | 35 | 36 | | | | |
| Médias (ºC) | | 60,86 | 60,79 | 10,41 | 10,18 | 60,86 | 9,93 | - | 9,77 | 60,41 | | | | |
| **Correlações do Nodo 18** | | X | X | | | X | | | | X | | | | |

A Tabela 9 ilustra as correlações dos nodos 4, 36, 8 e 18. O nodo

34 não possui valores da média (Tabela 9). Ele pode ter sofrido alguma interferência na comunicação pela rede sem fio e não conseguiu entregar o seu pacote. Em RSSFs esse é um problema comum, e graças à grande redundância de nodos, a área monitorada não ficou comprometida pela falta do nodo 34.

O segundo evento foi detectado com sucesso também. O fogo iniciou em algum ponto entre os nodos 26 e 39. Esses foram os primeiros a terem as médias elevadas. Na próxima iteração, o coordenador detectou alteração nas leituras nos nodos próximos ao evento.

## 5.5 CONSIDERAÇÕES DO CAPÍTULO

Neste capítulo foi apresentado o cenário para a realização das simulações necessárias para validar a proposta. Os resultados obtidos se mostraram promissores com relação à eficácia na detecção e identificação de *outliers*.

A utilização de técnicas de detecção baseadas em estatísticas possui muitas vantagens para RSSFs, como o baixo custo computacional. Entre os métodos avaliados, os métodos CWA+FTA e FTA se destacaram na eliminação de todos os *outliers*. Portanto, as técnicas CWA+FTA e FTA alcançaram bons resultados nos cenários simulados de RSSFs de larga escala conferindo boa detecção, precisão e confiabilidade no monitoramento.

O método de Peirce não detecta todos os *outliers* conforme mostrado na primeira parte da avaliação da abordagem. Entretanto, ele pode ser usado para identificar eventos aliando-se à correlação espacial e limites pré-definidos, conforme demonstrado na segunda parte da avaliação da proposta.

Por fim, a proposta avaliada para este cenário detectou e identificou os *outliers* e eventos inseridos. Foram detectados os *outliers* e posteriormente identificadas as possíveis correlações entre eles, resultando na identificação dos eventos.

# 6 CONSIDERAÇÕES FINAIS

Neste último capítulo são revistas as motivações e objetivos do trabalho. Uma breve revisão do que foi proposto e desenvolvido e dos resultados alcançados são apresentados na seção visão geral do trabalho. As limitações encontradas, como também os outros caminhos que poderiam ser seguidos são discutidos na seção de trabalhos futuros.

## 6.1 REVISÃO DAS MOTIVAÇÕES DO TRABALHO

A utilização das RSSFs se expande em várias áreas e aplicações do nosso dia-a-dia, sendo uma importante área de pesquisa. Estas redes estão mais onipresentes nas nossas vidas com o desenvolvimento dos sensores inteligentes. A confiabilidade dos dados gerados por estas redes é uma questão de grande relevância, pois decisões são tomadas a partir dos dados coletados pelas RSSFs.

Um problema recorrente no monitoramento de grandes áreas utilizando RSSFs é o grande volume de dados gerados e a falta de confiabilidade dos dados. Devido à utilização de uma grande quantidade de nodos de baixo custo e limitação de recursos, *outliers* podem estar contidos nos conjuntos de leituras sensoriadas. Esses *outliers* podem ser resultantes de dados espúrios ou de eventos.

As restrições computacionais de recursos das RSSFs delimitam as técnicas que podem ser empregadas para detecção e identificação de *outliers*. Portanto, este trabalho teve como principal objetivo analisar as técnicas de baixo custo computacional que permitam detectar e identificar *outliers* em RSSFs de larga escala, através do uso de técnicas baseadas em estatística em nodos de baixo custo e com restrições de *hardware*.

## 6.2 VISÃO GERAL DO TRABALHO

Este trabalho teve como objetivo realizar a detecção de dados anômalos (*outliers*) e, posteriormente, fazer a distinção entre dados espúrios e eventos. Para tal, uma abordagem para realizar a detecção e identificação de *outliers* foi proposta. A abordagem foi dividida em duas etapas, a primeira é a detecção dos *outliers*, com o processo de detecção ocorrendo localmente no nodo. A segunda etapa realiza a

identificação dos *outliers* detectados na etapa anterior, utilizando correlações espaciais e limites pré-definidos executados no coordenador.

A RSL foi utilizada na seleção de trabalhos correlatos, principalmente porque esse método possui um protocolo de busca que facilita a localização de todos os trabalhos pertinentes em uma área específica. Na RSL foram selecionados 19 artigos que representam o estado da arte atual relacionado com o tema. Foram encontradas lacunas entre as abordagens, como a baixa discussão das técnicas aplicadas em RSSFs larga escala.

A utilização do simulador de RSSFs OMNeT++/Castalia, permitiu a elaboração de um cenário de monitoramento de larga escala a partir da aplicação real desenvolvida pelo projeto Sensorscope. O simulador ainda permitiu a implementação dos algoritmos referentes à abordagem.

A avaliação da abordagem também ocorreu em duas etapas. Primeiro, foram escolhidas as técnicas baseadas em estatística para implementar no processo de detecção. As técnicas avaliadas foram: Chauvenet, Peirce, FTA e CWA+FTA. Após os testes e execuções no cenário de simulação, foi concluído que a técnica CWA+FTA era a melhor para o cenário aplicado.

Na simulação da segunda parte da abordagem foi avaliada a utilização da correlação entre os nodos com a combinação de limites pré-definidos, com o intuito de identificar possíveis eventos ou dados espúrios, provenientes das médias realizadas no processo de detecção enviado ao coordenador. O método proposto detectou corretamente todos os eventos simulados, não interpretando nenhum como falso alarme, e também identificou corretamente os nodos com dados espúrios.

A escolha das técnicas baseadas em estatísticas foi motivada pelas restrições intrínsecas das RSSFs, e por serem técnicas de baixo custo computacional, baixa complexidade de implementação e adaptabilidade ao cenário sem precisar de conhecimento prévio dos dados.

Este trabalho contribuiu para mostrar que a utilização das técnicas baseadas em estatísticas em RSSFs de larga escala permitiu alcançar os objetivos de detectar e identificar *outliers*.

## 6.3 TRABALHOS FUTUROS

Devidos às limitações delimitadas pelo escopo do trabalho, algumas questões julgadas importantes acabaram não sendo contempladas.

Abaixo seguem algumas ideias e melhorias que ficam como temas

para trabalhos futuros:

- Ampliar o cenário e quantidade de nodos para avaliar o comportamento dos métodos baseadas em estatísticas na detecção de *outliers* em RSSFs;
- Inserir perturbações na RSSF utilizada para estudar como as técnicas se comportaram com interferência, aumento das perdas de pacotes entre outros;
- Modificar os tempos de leituras dos sensores para minimizar os tempos de resposta do processo de identificação;
- Expandir a abordagem para uma arquitetura. Propor uma arquitetura para manipulação dos dados coletados na RSSF, que envolva as etapas de detecção, identificação e tratamento dos dados. Os dados seriam multivariados, ampliando para um cenário multidimensional, além de possibilitar a correlação espacial-temporal. A correlação entre atributos também seria levada em consideração na identificação dos *outliers*. A arquitetura ainda proveria técnicas para identificar ataques maliciosos na RSSF, além de identificar a localização dos nodos que não possuem sistemas de localização geográfica. O modo de detecção seria oferecido localmente e distribuído.

# REFERÊNCIAS

ABAFIRE. **2.2) Detectores / Sensores de Temperatura, Térmicos e Termovelocimétricos - Convencionais**. 2017. Disponível em: <http://abafire.com.br/categorias/detectores-sensores-de-temperatura-termicos-e-termovelocimetricos-convencionais/>.

AKYILDIZ, I. F. et al. A survey on sensor networks. **IEEE Communications Magazine**, v. 40, n. 8, p. 102–105, 2002.

AKYILDIZ, I. F. et al. Wireless Sensor Networks: A Survey. **Computer Networks**, Elsevier North-Holland, Inc., New York, NY, USA, v. 38, n. 4, p. 393–422, 2002.

AMIDI, A.; HAMM, N. A. S.; MERATNIA, N. Wireless sensor networks and fusion of contextual information for weather outlier detection. In: **International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives**. Tehran, Iran: University of Twente in the Netherlands, 2013. v. 40, n. 1W3, p. 37–41.

ANDRADE, A. **Abordagem para Detecção, Identificação e Tratamento de Outliers em Rede de Sensores Sem Fio de Larga Escala**. Florianópolis: Qualificação de doutorado - Universidade Federal de Santa Catarina, 2016. 88 p.

ATZORI, L.; IERA, A.; MORABITO, G. The Internet of Things : A survey. **Computer Networks**, Elsevier B.V., v. 54, n. 15, p. 2787–2805, 2010.

BAHREPOUR, M. et al. Use of event detection approaches for outlier detection in wireless sensor networks. In: **Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP)**. Melbourne, Australia: 2009 5th International Conference on. IEEE, 2009. p. 439–444.

BARNETT, V.; LEWIS, T. **Outliers in Statistical Data**. 3. ed. New York: Wiley, 1994. (Wiley Series in Probability & Statistics).

BASSI, A.; HORN, G. Internet of things in 2020: A roadmap for the future. **European Commission: Information Society and Media**, v. 22, p. 97–114, 2008.

BHOJANNAWAR, S. S.; BULLA, C. M.; DANAWADE, V. M. Anomaly Detection Techniques for Wireless Sensor Networks - A Survey. **International Journal of Advanced Research in Computer and Communication Engineering**, v. 2, n. 10, p. 3852–3857, 2013.

BOSMAN, H. H. et al. Spatial anomaly detection in sensor networks using neighborhood information. **Information Fusion**, Elsevier B.V., v. 33, p. 41–56, 2017.

BOULIS, A. **Castalia A simulator for Wireless Sensor Networks and Body Area Networks - User's Manual**. digital, 2010. 79 p. Disponível em: <http:/ /castalia.npc.nicta.com.au/%5Cnhttp://castalia.npc.nicta.com.au/pdfs/Castalia - User Manual.pdf>.

CALLEGARO, R. **Uma Arquitetura para Fusão de Dados e Detecção de Outliers em Sensores de Baixo Custo de Redes de Sensores sem Fio**. 130 p. Tese (Doutorado) — Universidade Federal de Santa Catarina, 2014.

CAO, S. S. b. et al. Anomaly event detection method based on compressive sensing and iteration in wireless sensor networks. **Journal of Networks**, v. 9, n. 3, p. 711–718, mar 2014.

CERPA, A. et al. Habitat monitoring: Application driver for wireless communications technology. **ACM SIGCOMM Computer Communication Review**, ACM, v. 31, n. 2 supplement, p. 20–41, 2001.

CHANDOLA, V.; BANERJEE, A.; KUMAR, V. Outlier Detection : A Survey. **ACM Computing Surveys**, v. 41, n. 3, p. 241, 2009.

CHENG, P.; ZHU, M. Lightweight Anomaly Detection for Wireless Sensor Networks. **International Journal of Distributed Sensor Networks**, Hindawi Publishing Corporation, v. 2015, n. 1, 2015.

CHOUIKHI, S. et al. A survey on fault tolerance in small and large scale wireless sensor networks. **Computer Communications**, Elsevier Ltd., v. 69, p. 22–37, 2015. ISSN 01403664. Disponível em: <http://dx.doi.org/10.1016/j.comcom.2015.05.007>.

DARGIE, W.; POELLABAUER, C. **Fundamentals of Wireless Sensor Networks: Theory and Practice**. [S.l.]: John Wiley & Sons Ltda, 2010. 330 p.

ELMENREICH, W. Fusion of Continuous-valued Sensor Measurements using Confidence-weighted Averaging. **Journal of Vibration and Control**, Sage Science Press (Sage Publications), Thousand Oaks, CA, v. 13, n. 9-10, p. 1303–1312, 2007.

FAWZY, A.; MOKHTAR, H. M. O.; HEGAZY, O. Outliers detection and classification in wireless sensor networks. **Egyptian Informatics Journal**, Ministry of Higher Education and Scientific Research, v. 14, n. 2, p. 157–164, 2013.

GIL, P.; SANTOS, A.; CARDOSO, A. Dealing with outliers in wireless sensor networks: An oil refinery application. **IEEE Transactions on Control Systems Technology**, v. 22, n. 4, p. 1589–1596, 2014.

GIUSTO, D. et al. **The Internet of Things: 20th Tyrrhenian Workshop on Digital Communications**. New York: Springer Science & Business Media, 2010. 452 p.

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

GUGLIELMO, D. D.; BRIENZA, S.; ANASTASI, G. IEEE 802.15.4e: A survey. **Computer Communications**, v. 88, p. –, 2016.

HAWKINS, D. **Identification of Outliers**. London: Chapman and Hall, 1982. 590–597 p.

HOGG, R. V.; TANIS, E.; ZIMMERMAN, D. **Probability and Statistical Inference**. 9. ed. [S.l.]: Pearson Higher, 2003. 552 p.

IEEE. **IEEE Standard for Low-Rate Wireless Networks**. New York: The Institute of Electrical and Electronics Engineers, 2015.

ILYAS, M.; MAHGOUB, I. **Handbook of Sensor Networks: Compact Wireless and Wired Sensing Systems**. New York: CRC PRESS, 2004.

INGELREST, F. et al. SensorScope: Application-Specific Sensor Network for Environmental Monitoring. **ACM Transactions on Sensor Networks**, v. 6, n. 2, p. 1–32, 2010.

JAYASHREE, L. S.; ARUMUGAM, S.; VIJAYALAKSHMI, K. A robust outlier detection scheme for collaborative sensor networks. **Journal of Digital Information Management**, v. 5, n. 1, p. 12–18, 2007.

JURDAK, R. et al. Chapter 12 Wireless Sensor Network Anomalies: Diagnosis and Detection Strategies. In: **Intelligent-Based Systems Engineering**. Berlin: Springer Berlin Heidelberg, 2011. p. 309–325.

KALAYCI, T. E. et al. How wireless sensor networks can benefit from Brain Emotional learning Based Intelligent Controller (BELBIC). In: **Procedia Computer Science**. [S.l.]: Elsevier B.V., 2011. v. 5, p. 216–223.

KITCHENHAM, B. **Procedures for performing systematic reviews**. Eversleigh, 2004. v. 33, n. TR/SE-0401, 28 p.

LEÓN, O.; HERNÁNDEZ-SERRANO, J.; SORIANO, M. Securing cognitive radio networks. **International Journal of Communication Systems**, v. 23, n. 5, p. 633–652, 2010.

LOUREIRO, A. a.F. et al. Redes de Sensores Sem Fio. **XXI Simpósio Brasileiro de Redes de Computadores**, p. 179–226, 2003.

MARZULLO, K. Tolerating failures of continuous-valued sensors. **ACM Transactions on Computer Systems**, v. 8, n. 4, p. 284–304, 1990.

MIORANDI, D. et al. Ad Hoc Networks Internet of things: Vision , applications and research challenges. **Ad Hoc Networks**, Elsevier B.V., v. 10, n. 7, p. 1497–1516, 2012.

MOSHTAGHI, M. et al. An adaptive elliptical anomaly detection model for wireless sensor networks. **Computer Networks**, Elsevier B.V., v. 64, p. 195–207, 2014.

NAKAMURA, E. F.; LOUREIRO, A. a. F.; FRERY, A. C. Information fusion for wireless sensor networks. **ACM Computing Surveys**, v. 39, n. 3, p. 9–es, 2007.

NBR ISO 7240-5. **Sistemas de detecção e alarme de incêndio Parte 5: Detectores de temperatura pontuais**. 2008.

OLIVEIRA, L. M. L. Wireless Sensor Networks : a Survey on Environmental Monitoring. **Journal of commnications**, v. 6, n. 2, p. 143–151, 2011.

PANTELAKI, K.; PANAGIOTAKIS, S.; VLISSIDIS, A. Survey of the IEEE 802.15.4 Standard ' s Developments for Wireless Sensor Networking. **American Journal of Mobile Systems, Applications and Services**, v. 2, n. 1, p. 13–31, 2016.

PEI, X. et al. Spatio-temporal Event Detection: A Hierarchy Based Approach for Wireless Sensor Network. In: **2014 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery**. Washington: IEEE Computer Society, 2014. p. 372–379.

RAJASEGARAR, S. et al. Elliptical anomalies in wireless sensor networks. **ACM Transactions on Sensor Networks**, v. 6, n. 1, p. 1–28, 2009.

RAJASEGARAR, S. et al. Ellipsoidal neighbourhood outlier factor for distributed anomaly detection in resource constrained networks. **Pattern Recognition**, Elsevier, v. 47, n. 9, p. 2867–2879, 2014.

RAJASEGARAR, S.; LECKIE, C.; PALANISWAMI, M. Anomaly detection in wireless sensor networks. **IEEE Wireless Communications**, v. 15, n. 4, p. 34–40, 2008.

RASSAM, M. A.; MAAROF, M. A.; ZAINAL, A. Adaptive and online data anomaly detection for wireless sensor systems. **Knowledge-Based Systems**, Elsevier B.V., v. 60, p. 44–57, 2014.

RASSAM, M. A.; ZAINAL, A.; MAAROF, M. A. Advancements of data anomaly detection research in Wireless Sensor Networks: A survey and open issues. **Sensors**, v. 13, n. 8, p. 10087–10122, 2013.

ROSS, S. M. Peirce's Criterion for the Elimination of Suspect Experimental Data. **Journal of Engineering Technology**, v. 20, n. 2, p. 1–12, 2003.

RUIZ, L. B. et al. Arquiteturas para Redes de Sensores Sem Fio. In: **22 Simpósio Brasileiro de Redes de Computadores**. [S.l.: s.n.], 2004. p. 167–218.

RUIZ, M. et al. The Convergence between Wireless Sensor Networks and the Internet of Things; Challenges and Perspectives: a Survey.

**IEEE Latin America Transactions**, v. 14, n. 10, p. 4249–4254, 2016.

SALEM, O. et al. Anomaly Detection Scheme for Medical Wireless Sensor Networks. In: **Handbook of Medical and Healthcare Technologies**. New York, NY: Springer New York, 2013. p. 207–222.

SHAHID, N.; NAQVI, I. H.; Bin Qaisar, S. SVM Based Event Detection and Identification: Exploiting Temporal Attribute Correlations Using SensGru. **Mathematical Problems in Engineering**, Hindawi Publishing Corporation, v. 2014, p. 1–12, 2014.

SHAHID, N.; NAQVI, I. H.; QAISAR, S. B. Characteristics and classification of outlier detection techniques for wireless sensor networks in harsh environments: a survey. **Artificial Intelligence Review**, v. 43, n. 2, p. 193–228, feb 2015.

SHENG, B. et al. Outlier detection in sensor networks. In: **Proceedings of the 8th ACM international symposium on Mobile ad hoc networking and computing**. New York: ACM, 2007. p. 219–228.

Silicon Laboratories. **The Evolution of Wireless Sensor Networks**. 2013. 1–5 p.

SOARES, M. **Critério de Chauvenet**. 2013. Disponível em: <http://www.mspc.eng.br/tecdiv/med200.shtml#crit_chauvenet>.

TAYLOR, J. R. **Introdução à Análise de Erros: O Estudo de Incertezas em Medições Físicas**. 2. ed. [S.l.]: Bookman, 2012.

THE Mahalanobis distance. **Chemometrics and Intelligent Laboratory Systems**, v. 50, n. 1, p. 1 – 18, 2000.

VIEIRA, M. A. M. et al. Survey on wireless sensor network devices. In: IEEE. **Emerging Technologies and Factory Automation, 2003. Proceedings. ETFA'03. IEEE Conference**. [S.l.], 2003. v. 1, p. 537–544.

WANG, C.; LIN, H.; JIANG, H. Trajectory-based multi-dimensional outlier detection in wireless sensor networks using Hidden Markov Models. **Wireless Networks**, v. 20, n. 8, p. 2409–2418, nov 2014.

WANG, H. et al. Connectivity, coverage and power consumption in large-scale wireless sensor networks. **Computer Networks**, v. 75, p. 212 – 225, 2014. ISSN 1389-1286. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1389128614003636>.

WANG, Y. et al. Efficient event detection using self-learning threshold for wireless sensor networks. **Wireless Networks**, v. 21, n. 6, p. 1783–1799, aug 2015.

WU, W. et al. Localized outlying and boundary data detection in sensor networks. **IEEE Transactions on Knowledge and Data Engineering**, v. 19, n. 8, p. 1145–1156, 2007.

XIE, M.; HU, J.; GUO, S. Segment-based anomaly detection with approximated sample covariance matrix in wireless sensor networks. **IEEE Transactions on Parallel and Distributed Systems**, v. 26, n. 2, p. 574–583, 2015.

YICK, J.; MUKHERJEE, B.; GHOSAL, D. Wireless sensor network survey. **Computer Networks**, v. 52, n. 12, p. 2292–2330, 2008.

YIN, J.; HU, D.; YANG, Q. Spatio-Temporal Event Detection Using Dynamic Conditional Random Fields. In: **Proceedings of the Twenty-First International Joint Conference on Artificial Intelligence**. New York: International Joint Conferences on Artificial Intelligence, 2009. p. 1321–1326.

ZHANG, P. et al. Hardware design experiences in zebranet. In: ACM. **Proceedings of the 2nd international conference on Embedded networked sensor systems**. [S.l.], 2004. p. 227–238.

ZHANG, R. et al. Cooperative sensor anomaly detection using global information. **Tsinghua Science and Technology**, v. 18, n. 3, p. 209–219, 2013.

ZHANG, Y. **Observing the Unobservable Distributed Online Outlier Detection in Wireless Sensor Networks**. 174 p. Tese (Doutorado) — University of Twente, 2010.

ZHANG, Y.; MERATNIA, N.; HAVINGA, P. Outlier Detection Techniques for Wireless Sensor Networks: A Survey. **IEEE Communications Surveys & Tutorials**, IEEE Press, Piscataway, NJ, USA, v. 12, n. 2, p. 159–170, 2010.

# APÊNDICE A – Revisão Sistemática da Literatura: Protocolo

Este apêndice é dedicado a apresentação da Revisão Sistemática da Literatura (RSL) realizada neste trabalho. A RSL é composta por etapas e um protocolo para a realização e documentação dessas etapas.

Segundo Kitchenham (2004), RSL significa identificar, avaliar e interpretar todas as pesquisas relevantes disponíveis para um determinado assunto, área ou fenômeno de interesse.

Abaixo apresentam-se todas as etapas realizadas para a elaboração da RSL. Foi utilizada a ferramenta StArt[1] (*State of the Art through Systematic Review*) desenvolvida pela Universidade Federal de São Carlos para conduzir a RSL.

## A.1 OBJETIVO

Levantar o Estado da Arte em relação as técnicas de detecção e identificação de *outlier* em Redes de Sensores sem Fio de Larga Escala.

## A.2 QUESTÃO PRINCIPAL

Quais são as técnicas utilizadas para detectar e identificar *outlier* em RSSF de Larga Escala com abordagem local.

- **População**: Artigos ou estudos que utilizam técnicas de detecção e identificação de *outlier* em RSSF de larga escala.
- **Intervenção**: Técnicas de detecção e identificação de *outlier*.
- **Controle**: Estudar as técnicas de detecção e identificação de *outlier* em RSSFs.
- **Resultados**: Obtenção do Estado da Arte e lacunas que possam ser trabalhadas.

## A.3 CRITÉRIOS DE BUSCA

### A.3.1 Palavras-Chave e Sinônimos

Foram utilizadas as seguintes palavras-chave e sinônimos: *outlier*, *anomaly*, *event*, *detection*, *classification*, *identification* e *wireless sensor network*.

[1] http://lapes.dc.ufscar.br/tools/start_tool

### A.3.2 Bases de Dados

Foram utilizadas bibliotecas digitais *online* e bases de dados indexadas. Cada base de dados possuía características individuais da estrutura da expressão de busca a ser inserida na busca avançada. Abaixo segue a lista de bases de dados utilizadas juntamente com sua expressão de busca.

- **Engineering Village**[2]: ((((outlier OR anomaly) AND event AND (detection OR classification OR identification)) WN AB) AND ((“wireless sensor networks”) WN All fields))
- **ACM**[3]: Searched for recordAbstract:((outlier OR anomaly) AND event AND (detection OR classification OR identification)) AND (+“wireless sensor networks”)
- **IEEE**[4]:(outlier OR anomaly) AND event AND (detection OR classification OR identification)) AND “wireless sensor networks”)
- **ScienceDirect**[5]: ABSTRACT((outlier OR anomaly) AND event AND (detection OR classification OR identification)) and (“wireless sensor networks”)
- **Scopus**[6]: (ABS((outlier OR anomaly) AND event AND (detection OR classification OR identification )) AND TITLE-ABS-KEY (“wireless sensor networks”))
- **Web of Science**[7]: (outlier OR anomaly) AND event AND (detection OR classification OR identification) AND “wireless sensor networks”

## A.4 CRITÉRIOS PARA SELEÇÃO INICIAL

Foram definidos os critérios de inclusão e exclusão que serão aplicados nos trabalhos resultantes da busca anterior.

[2] https://www.engineeringvillage.com/search/quick.url
[3] http://dl.acm.org/advsearch.cfm?coll=DL&dl=ACM&CFID=943752029&CFTOKEN=76069344
[4] http://ieeexplore.ieee.org/Xplore/home.jsp
[5] http://www.sciencedirect.com/
[6] https://www.scopus.com/
[7] https://www.webofknowledge.com/

### A.4.1 Critérios de Inclusão

- Trabalhos que abordem técnicas de detecção de *outlier*;
- Trabalhos que abordem técnicas de identificação de eventos;
- Trabalhos publicados em *journal* ou revista;
- Trabalhos disponíveis gratuitamente para *download*;
- Trabalhos disponíveis em uma das três línguas: inglês, português ou espanhol.

### A.4.2 Critérios de Exclusão

- Trabalhos que não abordem RSSF;
- Trabalhos que não usem uma abordagem local para detecção
- Trabalhos que não apresentem título e *abstract* pertinente a busca;
- Trabalhos com data de publicação inferior a 2000;
- Trabalhos duplicados, presentes em mais de uma base de dados.

## A.5 SELEÇÃO INICIAL DOS ESTUDOS

A seleção dos artigos é conduzida em duas etapas. A primeira com a leitura do título e *abstract*. A segunda parte é a leitura integral dos artigos.

## A.6 AVALIAÇÃO DE QUALIDADE DOS ESTUDOS

O artigo deve conter estes dados:

- O nome da técnica deve ser apresentada;
- A metodologia;
- Os resultados.

## A.7 EXTRAÇÃO DOS DADOS

A etapa extração dos dados é utilizada para armazenar informações rápidas sobre o trabalho lido. Esses dados de extração são informações importantes para que o pesquisador não precise reler todo o trabalho novamente. Os dados extraídos foram:

- Nome dos autores;
- Título do trabalho;
- Ano de publicação;
- palavras-chaves;
- Ruídos e erros;
- Eventos;
- Método utilizado;
- Descrição do funcionamento do método;
- Técnica de análise;
- Modelo de estrutura;
- Cenário;
- Dimensão dos dados;
- Modo de Operação;
- Correlação;
- Escalar;

## A.8 BUSCA

No total foram obtidos 336 artigos. Na Tabela 10 é possível observar a quantidade de artigo por bases de dados.

Tabela 10 – Quantidade de artigos por bases de dados.

| **Bases de Dados** | **Quantidade de Artigos** |
|---|---|
| Engineering Village | 81 |
| ACM | 9 |
| IEEE | 62 |
| ScienceDirect | 15 |
| Scopus | 103 |
| Web of Science | 66 |
| **Total** | **336** |

A primeira etapa para escolhas dos artigos inicia-se pela leitura do título e *abstract*, na Tabela 11 é apresentado a relação dos artigos resultante da duas fases. Na primeira fase foram aceitos 94 artigos, rejeitados 87 por não se encaixares aos critérios definidos, além da exclusão de 155 artigos duplicados.

A segunda fase é leitura integral dos artigos. Como resultados foram aceitos 19 artigos para compor o estado da arte.

Tabela 11 – Resultado da quantidade de artigos resultantes

| **Artigos** | **Primeira Fase** | **Segunda Fase** |
|---|---|---|
| Aceitos | 94 | 19 |
| Rejeitados | 87 | 75 |
| Duplicados | 155 | 0 |