# DETECÇÃO DE OUTLIERS MULTIVARIADOS EM REDES DE SENSORES

**Fabrício Geraldo Valadares**[1], **André Luiz Lins de Aquino**[2]
**Álvaro Rodrigues Pereira Jr**[1]

[1] Universidade Federal de Ouro Preto
Departamento de Computação
CEP 35400-000 – Ouro Preto – MG – Brasil

[2] Universidade Federal de Alagoas
Instituto de Computação
CEP 57072-970 – Maceió – AL – Brasil

fabricio@ppgcc.ufop.br, alla@ic.ufal.br, alvaro@iceb.ufop.br

***Resumo.*** *As redes de sensores coletam e processam dados que são, em muitos casos, multivariados. Contudo, devido à fragilidade dos dispositivos sensores, algumas medidas podem desviar do padrão normal dos dados monitorados. Estes desvios são conhecidos como outliers. Neste artigo apresentamos uma comparação entre três algoritmos para a detecção de outliers em dados multivariados. Estes algoritmos serão aplicados às redes de sensores sem fio. Nosso objetivo é permitir uma análise robusta sobre o conjunto de dados coletado por essas redes.*
***Palavras Chave****: Redes de sensores. Detecção de outliers. Dados multivariados.*

***Abstract.*** *Wireless sensor networks collect and process environmental data which are, in many cases, multivariate. However, due to the fragility of sensor devices, some measurements may deviate from the standard set of monitored data. These deviations are known as outliers. This article presents a comparison among three algorithms for detecting outliers in multivariate data applied to wireless sensor networks. Our goal is to allow a robust analysis on a sensor data set.*
***Keywords****: Wireless sensor network. Outliers detection. Multivariate data.*

## 1. Introdução

O mundo ao nosso redor possui uma variedade de fenômenos como temperatura, pressão e umidade, que podem ser monitorados por dispositivos com capacidade de sensoriamento, processamento e comunicação. Trabalhando cooperativamente, estes dispositivos são conhecidos como redes de sensores sem fio (RSSFs) [Akyildiz et al. (2002)]. A principal tarefa desta rede é o envio de informações ao observador externo, porém, devido às restrições dos nós sensores, estas informações podem conter erros. Tais informações podem representar fenômenos univariados ou multivariados.

Na literatura, erros ou anomalias em dados amostrados são conhecidos como *outliers*. Em redes de sensores, nós podemos definir os *outliers* como um conjunto de valores que desviam significativamente do padrão normal dos dados sensoriados [Zhang et al. (2010)]. De forma geral, a identificação de *outliers* é utilizada em aprendizagem de máquina, mineração de dados, estatística, teoria da informação e no pré processamento de dados para a detecção de intrusão e fraude [Chandola et al. (2009)]. Contudo, apenas recentemente estas técnicas tem atraído a atenção de pesquisas da área de rede de sensores [Zhang et al. (2010)].

Neste trabalho, utilizamos métodos gerais para o reconhecimento de *outliers* em dados multivariados: *Minimum Volume Ellipsoid* (MVE), *Minimum Covariance Determinant* (MCD), and *Max-Eigen Difference* (MED). Nosso objetivo é efetuar análises robustas sobre os dados coletados por uma rede de sensores. A princípio, caracterizamos o problema de detecção de *outliers* em redes de sensores para permitir que os projetistas das redes utilizem tal caracterização na escolha das melhores práticas para a detecção de *outliers* em aplicações gerais. Em seguida, realizamos, via simulação, uma comparação entre os três métodos, considerando cenários específicos para uma rede de sensores. As simulações são executadas em cada nó sensor, uma vez que estamos interessados em *outliers* locais. Elas demostraram que o MVE e o MCD podem melhorar significativamente a representatividade dos dados.

O restante do artigo é organizado como a seguir. Seção 2 apresenta os trabalhos relacionados. A caracterização do problema é apresentada na seção 3. Os métodos utilizados são descritos na seção 4. Seção 5 discorre sobre a avaliação da representatividade dos dados. Os resultados de nossos experimentos são apresentados na seção 6. Finalmente, a seção 7 apresenta as conclusões e futuros direcionamentos da pesquisa.

## 2. Trabalhos Relacionados

O trabalho de Zhang *et al.*[Zhang et al. (2010)], propõe um arcabouço para a detecção de *outliers* em redes de sensores. Os autores classificam a detecção em métodos baseados em: estatística, classificação, vizinho mais próximo, agrupamento e decomposição espectral. Ainda de acordo com Zhang, a maior parte dos estudos existentes não leva em consideração os dados multivariados, e assume que os dados sensoriados são univariados ou bivariados. Além disso, estes trabalhos consideram a correlação espacial e temporal entre os dados dos nós vizinhos, e desprezam a correlação entre os dados coletados por cada sensor, o que eleva a complexidade computacional.

Considerando a detecção de *outliers* em dados univariados e bivariados, o trabalho de Sheng [Sheng et al. (2007)] propõe o uso de métodos baseados em histogramas para remover anomalias em conjuntos de dados gerados por nós sensores. O método apresentado

é capaz de filtrar os não *outliers*, e classificar possíveis anomalias em conjuntos de dados. Mas, seu objetivo é a redução do custo de comunicação, e nenhuma consideração sobre a representatividade dos dados é realizada.

O trabalho de Bahrepour *et al.* [Bahrepour et al. (2009)] apresenta o uso de métodos baseados no reconhecimento de eventos para a identificação de *outliers* em redes de sensores. Ele utiliza conjuntos de dados com duas dimensões, e as simulações são executados considerando dados reais e simulados. O método apresentado é capaz de reconhecer eventos (um tipo de *outlier*) e também anomalias. Contudo, análises com mais de duas variáveis não foram realizadas.

O trabalho de Rajasegarar *et al.* [Rajasegarar et al. (2010)] considera o uso de dados multivariados. Em seu trabalho são apresentados três métodos elípticos para o reconhecimento de *outliers* em dados gerados por redes de sensores. De acordo com o autor, estes métodos mantém a mesma precisão de métodos centralizados, mas reduz o custo energético, uma vez que os cálculos são executados de maneira distribuída. Neste trabalho cada nó calcula a média e matriz de covariância de seus dados, em seguida envia estas informações para o sorvedouro, que irá definir estimadores globais para a locação e escala. Estes estimadores são enviados de volta aos nós, e sobre eles serão definidos os *outliers*. O autor considera *outliers* globais, calculados com os estimadores locais, e locais, calculados com estimadores locais.

Considerando esse panorama geral, em nosso trabalho, realizamos a detecção de *outliers* locais, e exploramos a correlação entre os dados do sensor. Um aspecto em destaque no nosso trabalho é a garantia da representatividade do conjunto de dados após a remoção das anomalias, uma vez que os demais trabalhos não se preocupam com tal análise.

## 3. Caracterização do Problema

O diagrama apresentado na figura 1 representa apropriadamente o ambiente de monitoramento de uma rede de sensores sem fio, onde é considerado o uso de métodos para a detecção de *outliers*. Neste diagrama, $\mathcal{N}$ representa o ambiente e o processo à ser medido, $E$ representa seu domínio espaço temporal e as características topológicas da área monitorada, e $P$ representa o fenômeno multivariado de interesse.

$$\begin{array}{l} \mathcal{N} \mid E \xrightarrow{P} \mathcal{V}^* \xrightarrow{S|\mathcal{C}} \mathcal{V} \xrightarrow{\Psi} \mathcal{V}' \\ \qquad\qquad\quad \downarrow R \qquad\qquad\quad \downarrow R \\ \qquad\qquad\quad D \qquad\qquad\quad \widehat{D} \end{array}$$

**Figura 1. Representação da detecção de *outliers* em redes de sensores sem fio.**

Se considerarmos um cenário ideal, sem a presença de ruídos, teríamos um conjunto de dados ideal $\mathcal{V}^*$. Com isso, podemos tomar um decisões ideias ($D$), se aplicarmos um conjunto de regras ($R$), ao conjunto de dados ($\mathcal{V}^*$). Porém, devido à complexidade em que as redes de sensores sem fio operam, *e.g.*, interferências do ambiente e limitações de *hardware* este caso ideal é proibitivo.

Para monitorar este ambiente, devemos considerar as redes de sensores sem fio como um conjunto de $o$ nós observadores $S = (S^1_{1..p}, \ldots, S^o_{1..p})$, onde $p$ indica a quantidade

de variáveis que cada nó pode monitorar. Este conjunto de nós é depositado sobre a área de interesse e executa uma amostragem sobre $\mathcal{V}^*$, gerando um novo conjunto de dados $\mathcal{V}$. Como ilustração, assuma que a função $f$ é o fenômeno de interesse, *e.g.*, temperatura e umidade em um instante $t$, então teremos que $f(t) = (f_1, f_2)(t)$. Assim, cada observação do fenômeno $f$ gera um conjunto de dados multivariados $(s_1, \ldots, s_o)$, em que $s_i$ é um vetor de $p$ variáveis. Portanto, cada amostragem capturada pela rede será um conjunto de dados no formato $o \times p \times n$, onde $o$ indica o número de nós, $p$ representa o número de variáveis observadas e $n$ indica o número de amostras.

Uma vez que as redes de sensores não correspondem ao ambiente ideal, o conjunto de dados $\mathcal{V}$ pode conter ruídos, em nosso caso, chamados de *outliers*. No diagrama estes ruídos são representados por $\mathcal{C}$.

No diagrama ilustrado na figura 1, a detecção de *outliers* é representada por

$$\Psi : \mathbb{R}^{o \times p \times n} \rightarrow \mathbb{R}^{o \times p \times n'} \mid n' \leq n,$$

onde, $n$ indica o número de amostras contidas em $\mathcal{V}$ e $n'$ indica a quantidade de amostras resultantes ao desconsiderarmos os *outliers*. É importante destacar que $\Psi$ é aplicada em cada nó, o que resulta nas operações $1 \leq i \leq o : \Psi = (\Psi_1, \ldots, \Psi_o)$, implicando em $\Psi_i : \mathbb{R}^{p \times n} \rightarrow \mathbb{R}^{p \times n'} \mid n' \leq n$.

A aplicação de $\Psi$ permite uma análise robusta do conjunto de dados coletados por uma rede de sensores sem fio, uma vez que estamos interessados nas tomadas de decisões $\widehat{D}$ que devem ser tão próximas quanto possível de $D$ ao considerarmos as regras $R$.

## 4. Métodos para Detecção de *Outliers*

Conjuntos de dados, sejam eles grandes ou pequenos, podem conter elementos que não são consistentes com a distribuição do restante dos dados que compõe o conjunto, *e.g.*, pontos que desviam em uma ou mais variáveis, impedindo a modelagem estatística e a correta análise dos dados [de Santana Giroldo and Barroso (2008)]. Estas anomalias são chamadas *outliers*.

Existem diversas definições para os *outliers*, *e.g.*, "um *outlier* é uma observação, ou um conjunto de observações, que parecem ser inconsistentes quando comparados ao restante do conjunto" [Barnett and Lewis (1994)]. Em redes de sensores, podemos definir *outliers* como medidas que desviam significativamente do padrão normal dos dados sensoriados [Sheng et al. (2007)]. Estamos interessados na identificação destes desvios que podem invalidar ou impedir a tomada de decisões ($D$).

Um método clássico para o reconhecimento de *outliers* em dados multivariados é a distância de Mahalanobis:

$$\text{MD}_i = \sqrt{(\mathcal{V}_i - T(\mathcal{V}))C(\mathcal{V})^{-1}(\mathcal{V}_i - T(\mathcal{V}))^\top},$$

onde $\mathcal{V}$ representa o conjunto de dados que desejamos analisar, modelado como uma matriz de dimensões $p \times n$. $\mathcal{V}_i$ representa a $i$-ésima amostra do conjunto de entrada, $T$ é um vetor de média aritmética simples, onde existe uma média para cada variável. A matriz de covariância é representada por $C$ e sua dimensão é igual a $p \times p$.

Para uma distribuição Normal Multivariada, $\text{MD}_i^2$ tem aproximadamente uma distribuição qui-quadrado, com $p$ graus de liberdade ($\chi_p^2$). Então, podemos definir os *outli-*

*ers* como aquelas medidas que ultrapassam um determinado quantil da distribuição qui-quadrado. Em nosso caso, o quantil de $0.975$ será considerado [Filzmoser et al. (2005)].

Mesmo sendo utilizada para a detecção de *outliers*, a Distância de Mahalanobis é fortemente influenciada por eles. Isso ocorre devido à fragilidade dos estimadores de locação e dispersão utilizados, respectivamente a média e a matriz de covariância [Rousseeuw and Driessen (1999), Filzmoser et al. (2005)]. Logo, são necessários estimadores que sofram menor interferência das anomalias.

Giroldo e Barroso [de Santana Giroldo and Barroso (2008)], listam três métodos robustos para a identificação de *outliers* em conjuntos de dados multivariados. Estes métodos são o *Minimum Ellipsoid Volume* (MVE), o *Minimum Covariance Determinant* (MCD) e o *Max-Eigen Difference* (MED). Com exceção do MED, os outros métodos usam a distância de Mahalanobis para a detecção de *outliers*, mas, substituindo a média e a matriz de covariância por estimadores robustos (MVE, MCD).

### 4.1. Método Baseado no *Minimum Volume Ellipsoid*

Este método utiliza a distância de Mahalanobis, porém substituindo os estimadores de locação e dispersão por estimadores que sofrem menor interferência das anomalias. O estimador MVE pode ser definido como um par $(T, C)$, que substituem a média e a matriz de covariância por um vetor $T(\mathcal{V})$ de tamanho $p$, e $C(\mathcal{V})$, uma matriz positiva semi-definida de tamanho $p \times p$. O determinante da matriz é mínimo, sujeito a

$$\#\{i; (v_t - T)C^{-1}(v_t - T)^t \leq a^2\} \geq h,$$

onde $\#$ é o número de elementos no conjunto, $h = \lfloor (n + p + 1)/2 \rfloor$ e $a^2$ é uma constante, como $\chi^2_{p;0.50}$ considerando que a maior parte dos dados seguem uma distribuição Normal [Rousseeuw and Zomeren (1990)]. Esta metodologia segue um elipsoide de volume mínimo que cobre $h$ pontos, onde $n/2 \leq h < n$.

Os estimadores iniciais são a média e a matriz de covariância, a partir destes dados será traçado um elipsoide de volume mínimo que definirá os "pontos bons", ou seja, que pertencem a um intervalo de confiança definido, que no nosso caso será considerado um intervalo de $97,5\%$. Esses "pontos bons" são utilizados para os cálculos das estimativas finais dos parâmetros de localização e escala, respectivamente a sua média e matriz de covariância [de Santana Giroldo and Barroso (2008)]. Segundo Alameddine *et al.* [Alameddine et al. (2010)], o ponto de ruptura do MVE pode chegar a $50\%$, quando $n$ aumenta. Esse ponto representa a fração de *outliers* na amostra que pode tornar o estimador completamente tendencioso.

### 4.2. Método Baseado no *Minimum Covariance Determinant*

O MCD é frequentemente utilizado na prática, particularmente devido à rápida execução de seu algoritmo [Filzmoser et al. (2005)]. O objetivo dessa técnica é encontrar $h$ observações que tornem o determinante da matriz de covariância clássica mínimo. O estimador de localização é então a média destes $h$ pontos, enquanto o estimador de escala será sua matriz de covariância. Para manter um compromisso entre a eficiência e a robustez do método, o subconjunto é definido como $h \approx 0,75\, n$, onde $n$ indica o tamanho da amostra [Filzmoser et al. (2005)].

O valor aproximado do ponto de ruptura do MCD é dado por $(n - h)/n$, considerando $h \approx 0,75\, n$, seu valor será $25\%$. O cálculo desse estimador torna a Distância de

Mahalanobis robusta, para isso, inicialmente é realizado um ajuste para evitar que dados da cauda da distribuição sejam erroneamente classificados como *outliers*. Em seguida, faz-se um cálculo utilizando a Distância de Mahalanobis clássica e robusta (obtida por intermédio dos estimadores MCD) para assim, identificar os ruídos no conjunto de dados.

### 4.3. Método Baseado no *Max-Eigen Difference*

O MED, ao contrário das outras técnicas, não utiliza a distância de Mahalanobis para identificar os *outliers*. O reconhecimento de anomalias é realizado por intermédio dos autovetores e autovalores da matriz de covariância, aos quais são aplicmo mados à norma euclidiana [de Santana Giroldo and Barroso (2008)]. Primeiro o algoritmo calcula os autovalores e autovetores de todo o conjunto, em seguida, ele realiza o cálculo, desconsiderando-se a amostra atualmente analisada. Após esse procedimento é realizado o cálculo da distância de cada amostra em relação ao conjunto total. O resultado será padronizado, e as amostras com maior auto-diferença são considerados *outliers*.

## 5. Avaliação da Representatividade dos Dados

Nessa seção apresentaremos a metodologia utilizada para geração do conjunto de dados $\mathcal{V}$ e os métodos de avaliação da representatividade dos dados após a identificação dos *outliers*.

### 5.1. Metodologia

Os dados considerados na avaliação da representatividade são baseados em dados reais disponíveis por Albuquerque [Albuquerque (2007)]. Eles foram simulados considerando o seguinte processo,

$$\mathcal{V}^* \xrightarrow{S} \mathcal{V}^r \xrightarrow{\phi} \mathcal{V}^s \xrightarrow{\mathcal{C}} \mathcal{V} \, .$$

Os dados $\mathcal{V}^r$ correspondem às $19$ variáveis que são derivadas de observações de fenômenos ambientais reais, que incluem a concentração de poluentes *n-hexane*, *methylcyclopentane*, *toluene*, *p-xylene* e $1,3,5$-*trimethylbenzene*. Utilizamos $72$ amostras destes fenômenos que correspondem à média de quatro horas de observação. Devido ao processo de amostragem, via precipitação, não é possível obter a discretização completa do fenômeno, por essa razão o simulamos.

Na simulação dos dados a média e a estrutura de covariância obtidas de $\mathcal{V}^r$ foram mantidas para compor o conjunto $\mathcal{V}^s$, porém, assumindo três casos, as distribuições Normal, Skew-Normal e T-Student. Estas distribuições foram utilizadas para simular a imprecisão inerente dos dispositivos de sensoriamento, que, por sua vez, nem sempre representam de maneira satisfatória o fenômeno atual. A distribuição Skew-Normal tem um desvio igual a $0,5$, enquanto a distribuição T-Student, de cauda mais pesada, tem dois graus de liberdade. Em nosso caso, assumimos que o fenômeno monitorado tem comportamento Normal.

O número de dados simulados varia em cada intervalo de acordo com os seguintes fatores $\phi = \{10, 20, 30, 40, 50\}$, resultando em um número total de amostras por variável igual a $|\mathcal{V}^s| = \{720, 1.440, 2.160, 2.880, 3.600\}$. Para simular os dados com anomalias ($\mathcal{V}$) um conjunto de dados $\mathcal{C}$ similar a $\mathcal{V}^s$ é gerado, porém considerando o valor da média multiplicado por $2^{10}$. Com o auxílio de um conjunto de dados gerados a partir de uma

distribuição Bernoulli com probabilidade igual a $10\%$, as amostras que conterão ruídos são escolhidas em $\mathcal{V}^s$ e substituídas por amostras de $C$ gerando, assim, o conjunto de dados $\mathcal{V}$.

O método de Monte Carlo foi considerado em nossas avaliações. Este é um método estatístico utilizado em simulações estocásticas. A quantidade de replicações foi calculada de acordo com Jain [Jain (1991)],

$$\textit{replicações} = \left(\frac{100 z S_d}{p_c \overline{\mathcal{V}'}}\right)^2,$$

onde $z$ representa uma constante de valor $1,96$, $S_d$ é o desvio padrão encontrado nas dez primeiras simulações, $\overline{\mathcal{V}'}$ é a média e $p_c$ é a porcentagem da média desejada como desvio, em nosso caso, $5\%$. Após executar as primeiras simulações identificamos que para obter uma boa representatividade dos resultados seria necessário utilizar aproximadamente $600$ replicações independentes.

Neste trabalho, as simulações, métodos e análises foram implementadas no *software* estatístico $R$ [R Development Core Team (2011)], que possui excelentes propriedades numéricas, conforme descrito em Almiron *et al.* [Almiron et al. (2009)]. Especificamente, o MVE utiliza a função `cov.mve` do pacote *mvoutlier*, o MCD utiliza as funções *adjusted quantile plot* `aq.plot` e *distance-distance plot* `dd.plot`, desenvolvidas por Filzmoser *et al.* [Filzmoser et al. (2005)]. As duas funções empregam o estimador MCD para tornar robusta a distância de Mahalanobis, mas a primeira executa um ajuste para prevenir que dados da cauda da distribuição seja erroneamente classificados como *outlier*. A segunda define os *outliers* utilizando a distância de Mahalanobis clássica e robusta (obtida por intermédio dos estimadores MCD) para identificar os ruídos no conjunto de dados, conforme descrito na seção 4. Resultados das duas funções serão representados por MCD-AQ e MCD-DD, respectivamente.

### 5.2. Métodos de Avaliação

Utilizamos quatro métodos para avaliar os resultados das simulações, a contagem de *outliers*, o valor absoluto do erro relativo [Frery et al. (2008)], o teste de hipótese – *Analysis of Variance* (ANOVA) [Thomson (1993)] e a avaliação de Medidas de Tendência Central (MTC). Seguindo nossa representação para um rede de sensores em fio com detecção de *outliers*, cada método representa as regras ($R$) para a tomada de decisões ($\widehat{D}$).

Por intermédio da contagem de *outliers* podemos verificar se todos os ruídos inseridos foram encontrados e indicar quando algum dado real foi classificado como *outlier*, *i.e.*, falsos positivos. O valor absoluto do erro relativo realiza uma comparação entre a médias dos dados sem ruídos $\mathcal{V}^s$ e a média dos dados após a identificação e eliminação dos *outliers* $\mathcal{V}'$. O valor absoluto do erro relativo é

$$R_{vaer} = 100 \max_p \left(\frac{\overline{\mathcal{V}^s} - \overline{\mathcal{V}'}}{\overline{\mathcal{V}^s}}\right),$$

onde $\overline{\mathcal{V}^s}$ e $\overline{\mathcal{V}'}$ são respectivamente a média dos dados antes da contaminação e após o processo de remoção de *outliers*. O erro é calculado para cada variável monitorada ($p$), e apenas o maior valor encontrado será utilizado.

O teste ANOVA é usado para verificar se existem diferenças significativas entre as médias de $\mathcal{V}^s$ e $\mathcal{V}'$. Este teste estatístico é calculado por

$$R_{anova} = \lambda_B^2 / \lambda_W^2,$$

onde $\lambda_B^2$ é a variância entre os conjuntos e $\lambda_W^2$ representa a variância dentro dos conjuntos. Baseado neste cálculo, o $p$-valor é usado para determinar a aceitação ou rejeição da hipótese nula $H_0$. A aceitação da hipótese nula ($H_0$) é valida para valores acima de $0,05$.

A última análise considera três medidas de tendência central, a média aritmética, a mediana e a média truncada, onde consideramos a porcentagem igual a $10\%$. Os cálculos são realizados para cada variável, e apenas a maior diferença será utilizada. Assim obtemos as diferenças das medidas de tendência central entre os conjuntos $\mathcal{V}^s$ e $\mathcal{V}'$.

## 6. Resultados e Avaliações

Como mencionado na Seção 4, os resultados da distância de Mahalanobis ao quadrado ($\text{MD}^2$) são considerados *outliers* se seu valor exceder um certo quantil da distribuição qui-quadrado. Em nossas simulações utilizamos o padrão empregado pelos métodos, e o quantil foi definido como $0,975$. Assim, os resultados obtidos com métodos baseados nos estimadores MVE e MCD são considerados *outliers* se $\text{RD}^2 > \chi^2_p; 0,975$, onde $\text{RD}^2$ indica a distância de Mahalanobis Robusta. Os resultados para o método MED são considerados *outliers* se seu valor for maior do que a média dos valores encontrados.

Os parâmetros utilizados nas simulações são apresentados na tabela 1. Note que, considerando o número de variáveis ($p$) e o tamanho das amostras ($n$), $\mathcal{V}^s$ poderá ter os seguintes tamanhos $\{19 \times 720, 19 \times 1.440, 19 \times 2.160, 19 \times 2.880, 19 \times 3.600\}$. Além disso, como descrito na seção 5, os ruídos são inseridos com auxílio de dados gerados a partir de uma Distribuição de Bernoulli com probabilidade igual a $10\%$.

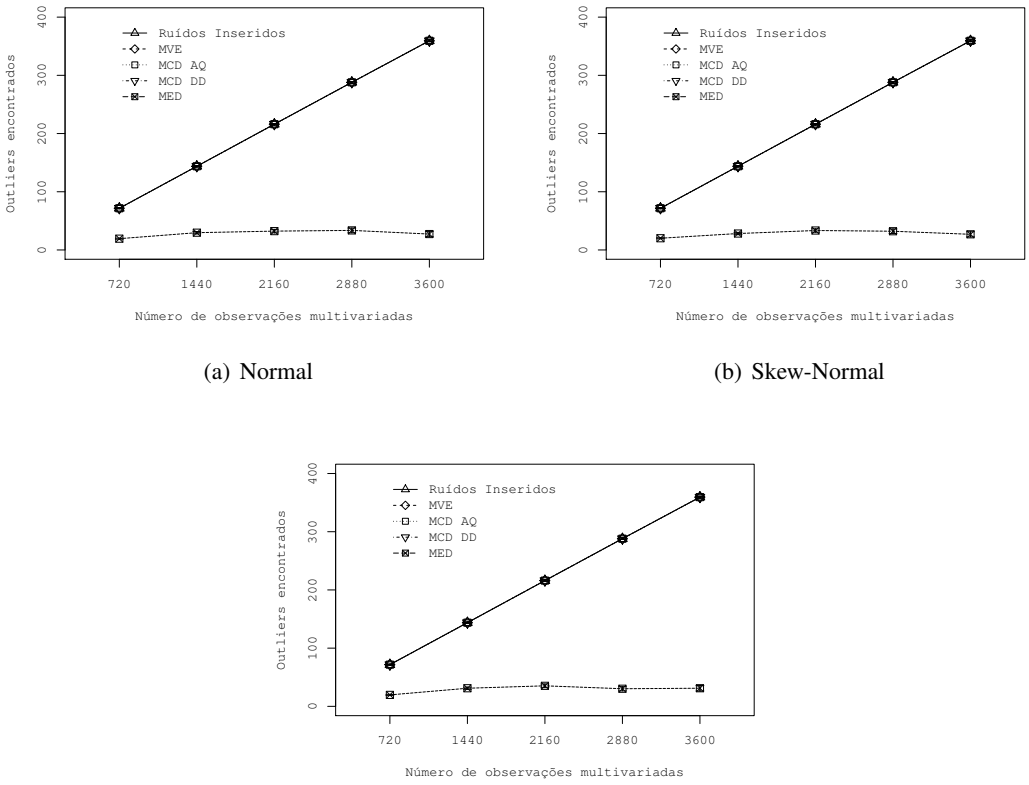
**Tabela 1. Parâmetros da simulação**

| Parâmetros | Valores |
|---|---|
| Número de variáveis (p) | 19 |
| Tamanho da amostra real | 72 |
| Amostras pseudo-reais (n) | 720, 1.440, 2.160, 2.880, 3.600 |
| Probabilidade de contaminação | 10.00% |
| Fator de multiplicação | $2^{10}$ |
| Replicações | 600 |

Nas avaliações iniciais, o número de falsos positivos estava elevado, então, inserimos um processo de verificação baseado na mediana (medida que sofre menor interferência das anomalias) e o desvio padrão. Assim, o MVE, MCD e MED classificam os candidatos a *outlier* e as amostras que tiverem valor superior à mediana $\pm$ desvio padrão são definitivamente classificados como *outliers*. O processo é simples, mas reduz significativamente o número de falsos positivos.

A primeira análise, ilustrada na figura 2, considera a contagem de *outliers* para cada método empregado, além da quantidade de ruídos inseridos. Nessa figura, o eixo $x$ representa o número de amostras, e o eixo $y$ a quantidade de ruídos encontrados. A curva do gráfico indica a quantidade de ruídos encontrados para cada tamanho de amostra considerado.

Na figura 2, As curvas do Ruído Inserido, MVE, MCD-AQ, e MCD-DD estão sobrepostas, em todas as distribuições consideradas. O que indica que todos os *outliers* inseridos foram localizados. Por outro lado, o MED encontrou apenas uma pequena quantidade de *outliers*. Isso ocorre devido à forma com que o MED classifica os *outliers*, ou seja, os autovalores e autovetores são calculados desconsiderando a amostra atualmente




(a) Normal

(b) Skew-Normal

(c) T-Student

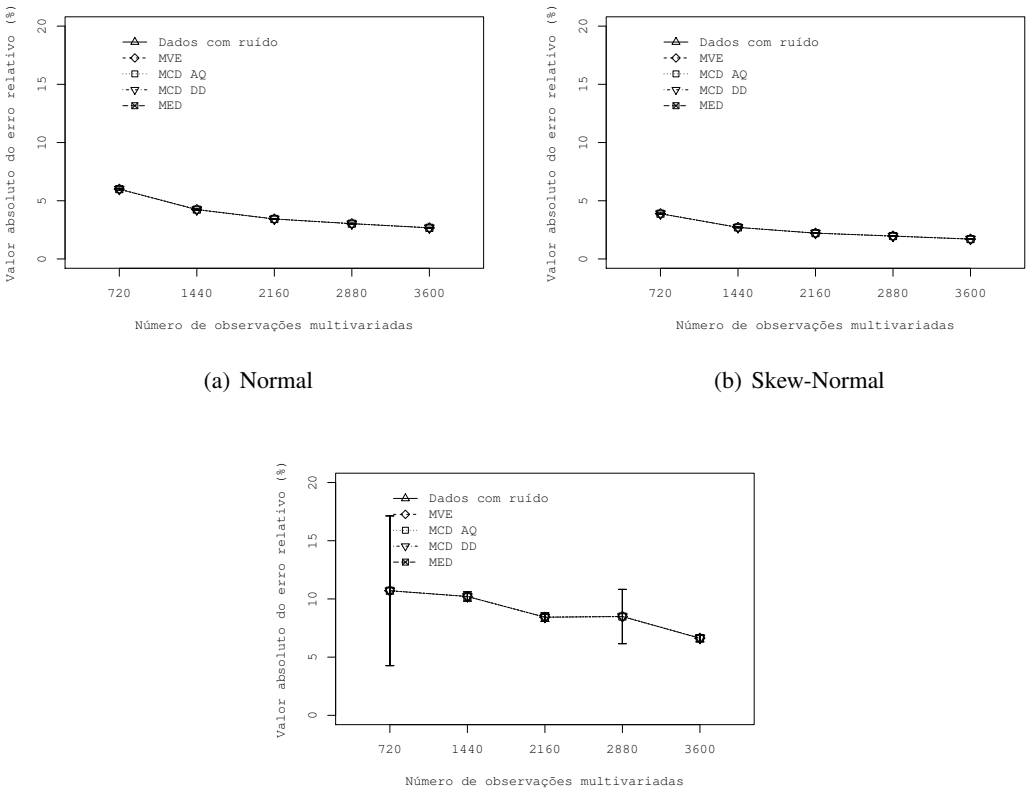
**Figura 2. Contagem de *outliers***

analisada, e como temos uma grande quantidade de dados, e com valores muito próximos, isso afeta o resultado.

Os resultados para $R_{vaer}$ são ilustrados na figura 3, onde o eixo $x$ indica o tamanho da amostra, e o eixo $y$ indica o valor absoluto do erro relativo. As curvas do MED não são exibidas porque elas excedem $50\%$ em muitas situações. Esses resultados inviabilizam a utilização do MED para as nossas aplicações.

Na figura 3, considerando as distribuições Normal e Skew-Normal, o valor máximo encontrado para o erro foi de $6\%$, com um conjunto de amostras de tamanho $720$. O melhor resultado foi encontrado em um conjunto de amostras de tamanho $3.600$, onde o erro é inferior a $3\%$. Na figura 3(c) estão os resultados da distribuição T-Student. O pior resultado foi observado em um conjunto de amostras com $720$ elementos, onde o erro foi de $17,2\%$. O melhor resultado foi encontrado em um conjunto de amostra com $3.600$ elementos, onde o erro é inferior a $6,7\%$. Estes resultados demonstram que as decisões $\widehat{D}$ baseadas na regra $R_{vaer}$ podem ser tomadas satisfatoriamente já que o erro demonstrado pode ser tolerado por grande parte das aplicações.

A tabela 2 apresenta os resultados para $R_{anova}$, utilizando a técnica MCD-AQ e todas as distribuições consideradas. Os resultados para o MVE e MCD-DD são similares, e foram omitidos. Como mencionado na seção 5.2, o $p$-valor acima de $0,05$ é válido para a aceitação da hipótese nula. Quando isso ocorre, o conjunto de dados após a remoção de




(a) Normal

(b) Skew-Normal

(c) T-Student

**Figura 3. Valor absoluto do erro relativo**

*outliers* representa de forma satisfatória o conjunto de dados antes da contaminação.

**Tabela 2. $R_{anova}$ com o MCD-AQ ($p$-valor)**

| Distribuição | $n = 720$ | $n = 1.440$ | $n = 2.160$ | $n = 2.880$ | $n = 3.600$ |
|---|---|---|---|---|---|
| Normal | $0{,}50 \pm 0{,}02$ | $0{,}48 \pm 0{,}02$ | $0{,}48 \pm 0{,}02$ | $0{,}48 \pm 0{,}02$ | $0{,}47 \pm 0{,}02$ |
| Skew-Normal | $0{,}55 \pm 0{,}02$ | $0{,}52 \pm 0{,}02$ | $0{,}52 \pm 0{,}02$ | $0{,}53 \pm 0{,}02$ | $0{,}52 \pm 0{,}02$ |
| T-Student | $0{,}14 \pm 0{,}01$ | $0{,}13 \pm 0{,}01$ | $0{,}10 \pm 0{,}01$ | $0{,}09 \pm 0{,}01$ | $0{,}07 \pm 0{,}01$ |

Conforme a tabela 2 a decisão $\widehat{D}$, baseada em $R_{anova}$ é valida para o MVE e para o MCD. Os resultados para o MED, mais uma vez, não foram satisfatórios. Como esta técnica não consegue identificar todos os *outliers*, sua utilização não traz benefícios à aplicação. Como o $p$-valor ficou abaixo de $0,05$ em todas as simulações, a hipótese nula é rejeitada para o MED.

Por fim, as mudanças nas medidas centrais são avaliadas, destacando que apenas a maior diferença entre $\mathcal{V}^s$ e $\mathcal{V}'$ será considerada. Também comparamos a diferença entre $\mathcal{V}^s$ e $\mathcal{V}$, ou seja, os dados sem *outliers* e os dados com *outliers*. Esta comparação é importante para entender como os ruídos podem interferir nas decisões que a aplicação pode tomar. A primeira avaliação é a média aritmética simples, pois ela é profundamente afetada na presença de *outliers*. Os resultados para as distribuições Normal, Skew-Normal e T-Student, utilizando o MVE e MCD são satisfatórios, e as diferenças encontradas foram inferiores a $0,6$. Quando consideramos o conjunto de dados na presença de *outliers*, a diferença entre

$\mathcal{V}^s$ e $\mathcal{V}$, é superior a $2.000$ unidades de medida. Os resultados do MED são bem próximos a este valor, já que ele não conseguiu remover todos os *outliers*.

A mediana e a média truncada, medidas que sofrem menor interferência dos *outliers*, obtiveram resultados satisfatórios considerando a aplicação do MVE e do MCD, para todos os cenários considerados. As diferenças encontradas para os dois estimadores não excedem $0,3$ unidades de medida, o que indica a robustez da técnica na presença de *outliers*. Como a mediana e a média truncada sofrem menor interferência dos *outliers*, as diferenças entre $\mathcal{V}^s$ e $\mathcal{V}$ são pequenas. Por exemplo, ao utilizarmos uma distribuição Normal, a diferença entre a mediana de $\mathcal{V}^s$ e $\mathcal{V}$, é igual a $2,23 \pm 0,03$, considerando um conjunto com $720$ amostras.

Os experimentos demonstraram que com a utilização do MVE e do MCD a representatividade dos dados pode ser mantida. Ou seja, não existem diferenças significativas entre o conjunto de dados $\mathcal{V}^s$ e $\mathcal{V}'$. Na presença de grande volumes de dados o MED demonstrou comportamento insatisfatório. Isso indica que a técnica não é apropriada para aplicações em redes de sensores que consideram os cenários aqui apresentados.

## 7. Conclusão

As redes de sensores não são utilizadas em ambientes controlados o que acarreta na geração de dados com erros, tais dados são considerados *outliers*. Estes *outliers* podem impedir ou invalidar a tomada de decisões por parte das aplicações. Neste artigo foram apresentadas três métodos para a identificação de *outliers* em dados multivariados, o MVE, o MCD e o MED. Estes métodos foram aplicados a cenários de aplicações de redes de sensores.

As avaliações indicaram que os resultados para o MVE e para o MCD são similares. Eles identificaram os *outliers* em todos os cenários considerados, e não apresentaram diferenças significativas entre o conjunto de dados sem *outliers* e o conjunto de dados após a remoção dos mesmos. O resultado dos testes permite dizer que as decisões tomadas sobre o conjunto de dados $\mathcal{V}'$ possuem uma boa representatividade quando comparadas às decisões tomadas sobre os dados originais. Em nossas simulações o MED não trouxe melhorias aos conjuntos analisados.

Como trabalho futuro, planejamos unir as técnicas aqui apresentadas com algoritmos para redução de dados. Estes algoritmos executam uma amostragem sobre conjuntos de dados multivariados, e seu objetivo é reduzir a quantidade de informação enviada até o sorvedouro, porém, mantendo a representatividade e correlação do conjunto original. Esperamos que a aplicação das técnicas aqui apresentadas torne robusta a redução de dados frente à presença de *outliers*. Neste trabalho consideramos *outliers* locais, mas em trabalhos futuros, os algoritmos apresentados podem ser utilizados para a detecção de *outliers* distribuídos, ou seja, considerando a combinação de dados de mais de um nó da rede.

## Referências

Akyildiz, I., Su, W., Sankarasubramaniam, Y., and Cayirci, E. (2002). A survey on sensor networks. *Communications Magazine, IEEE*, 40(8):102 – 114.

Alameddine, I., Kenney, M. A., Gosnell, R. J., and Reckhow, K. H. (2010). Robust multivariate outlier detection methods for environmental data. *Journal of Environmental Engineering*, 136(11):1299–1304.

Albuquerque, E. L. (2007). Compostos orgânicos voláteis na atmosfera urbana da região metropolitana de são paulo. Tese de Doutorado, Faculdade de Engenharia Química, Universidade Estadual de Campinas, Campinas, SP, Brasil.

Almiron, M. G., Almeida, E. S., and Miranda, M. N. (2009). The reliability of statistical functions in four software packages freely used in numerical computation. *Brazilian Journal of Probability and Statistics*, 23(2):107–119.

Bahrepour, M., Zhang, Y., Meratnia, N., and Havinga, P. (2009). Use of event detection approaches for outlier detection in wireless sensor networks. In *Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP), 2009 5th International Conference on*, pages 439 –444.

Barnett, V. and Lewis, T. (1994). *Outliers in statistical data*. Wiley series in probability and mathematical statistics: Applied probability and statistics. Wiley & Sons.

Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly detection: A survey. *ACM COMPUTING SURVEYS*, 41:15:1–15:58.

de Santana Giroldo, F. R. and Barroso, L. P. (2008). Alguns métodos robustos para detectar outliers multivariados. Master's thesis, Instituto de Matemática e Estatística - Universidade de São Paulo - IME-USP.

Filzmoser, P., Garrett, R. G., and Reimann, C. (2005). Multivariate outlier detection in exploration geochemistry. *Comput. Geosci.*, 31:579–587.

Frery, A. C., Ramos, H., Alencar-Neto, J., and Nakamura, E. (2008). Error estimation in wireless sensor networks. In *Proceedings of the 2008 ACM symposium on Applied computing*, SAC '08, pages 1923–1928, New York, NY, USA. ACM.

Jain, R. K. (1991). *The art of computer systems performance analysis: techniques for experimental design, measurement, simulation, and modeling*. Wiley professional computing. Wiley.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rajasegarar, S., Bezdek, J. C., Leckie, C., and Palaniswami, M. (2010). Elliptical anomalies in wireless sensor networks. *ACM Trans. Sen. Netw.*, 6:7:1–7:28.

Rousseeuw, P. J. and Driessen, K. V. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41:212–223.

Rousseeuw, P. J. and Zomeren, B. C. v. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, 85(411):pp. 633–639.

Sheng, B., Li, Q., Mao, W., and Jin, W. (2007). Outlier detection in sensor networks. In *Proceedings of the 8th ACM international symposium on Mobile ad hoc networking and computing*, MobiHoc '07, pages 219–228, New York, NY, USA. ACM.

Thomson, N. (1993). Understanding anova the apl way. In *Proceedings of the international conference on APL*, APL '93, pages 295–303, New York, NY, USA. ACM.

Zhang, Y., Meratnia, N., and Havinga, P. (2010). Outlier detection techniques for wireless sensor networks: A survey. *Communications Surveys Tutorials, IEEE*, 12(2):159 –170.