PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS
Programa de Pós-Graduação em Informática

REDUÇÃO DE DADOS MULTIVARIADOS EM REDES DE SENSORES SEM FIO

Orlando Silva Junior

Belo Horizonte
2009

Orlando Silva Junior

REDUÇÃO DE DADOS MULTIVARIADOS EM REDES DE SENSORES SEM FIO

Dissertação apresentada ao Programa de Pós-Graduação em Informática da Pontifícia Universidade Católica de Minas Gerais, como requisito parcial para a obtenção do título de Mestre em Informática.

Orientadora: Raquel Aparecida de Freitas Mini

Co-orientador: André Luiz Lins de Aquino

Belo Horizonte
2009

.

FICHA CATALOGRAFICA

Elaborada pela Biblioteca da Pontifícia Universidade Católica de Minas Gerais

S586r

Silva Junior, Orlando
Redução de dados multivariados em redes de sensores sem fio / Orlando Silva Junior. – Belo Horizonte, 2009.
100f. : il.

Orientadora: Raquel Aparecida de Freitas Mini.
Dissertação (Mestrado) – Pontifícia Universidade Católica de Minas Gerais. Programa de Pós-graduação em Informática.
Bibliografia.

1. Redes de computadores – Teses. 2. Sistemas de comunicação sem fios – Brasil. I. Mini, Raquel Aparecida de Freitas. II. Pontifícia Universidade Católica de Minas Gerais. III. Título

CDU: 681.3.01:621.39

Bibliotecário: Fernando A. Dias – CRB6/1084

PUC Minas
Programa de Pós-graduação em Informática

FOLHA DE APROVAÇÃO

"Redução de Dados Multivariados em Redes Sensores sem Fio"

Orlando Silva Júnior

Dissertação defendida e aprovada pela seguinte banca examinadora:

Profa. Raquel Aparecida Freitas Mini - Orientadora (PUC Minas)
Doutora em Ciências da Computação - UFMG

Prof. André Luiz Lins de Aquino - Co-orientador (UFOP)
Doutor em Ciência da Computação - UFMG

Prof. Luiz Henrique Andrade Correia (UFLA)
Doutor em Ciência da Computação - UFMG

Prof. Luis Enrique Zárate Gálvez - (PUC Minas)
Engenharia Metalúrgica e de Minas - UFMG

Belo Horizonte, 26 de Agosto de 2009.

Anel Rodoviário – KM 23,5, Rua Walter Ianni, 255 – São Gabriel / Belo Horizonte – MG / CEP: 31980-110
Fone: (31) 3439-5204 – Fax: (31) 3439-5204 – e-mail: mestradoinfsec@pucminas.br

AGRADECIMENTOS

A Deus, fonte de fé, força e esperança, por todas as graças a mim concedidas, principalmente pelo dom da vida e pela oportunidade de alcançar mais este objetivo.

Aos meus pais, aos quais devo tudo que sou e que tenho, pelo amor, carinho, amizade, e incentivo em todos os momentos da minha vida.

A minha amada Sabrina, pela presença constante, pelo amor, carinho, cumplicidade e apoio incondicionais.

A minha irmã Keli, pelo afeto e apoio sempre.

Aos meus amigos, em especial ao Sandro e ao Eurico, pela amizade, confiança e incentivo sempre.

Aos colegas do mestrado, especialmente à Bruna, pelo valioso auxílio na parte estatística, ao Alexander e ao Fernando, pelas monitorias.

À secretária Giovana, pela atenção e carinho com todos.

Aos colegas do sensornet, pelas valiosas sugestões.

Aos professores do mestrado da PUC Minas, pelo ensino de tão alta qualidade.

Por fim, aos grandes responsáveis pela realização desse trabalho. A minha orientadora Raquel, pelo comprometimento e pela experiência compartilhada. Ao grande idealizador desse projeto, meu também orientador André Lins, a quem devo grande parte desse sucesso e a quem tenho o grande orgulho de ter também como amigo, pessoa que esteve sempre presente, ajudando nos momentos mais complicados dessa jornada.

"O importante é estar pronto para, a qualquer momento, sacrificar o que somos pelo que podemos vir a ser."(Charles Du Bois)

RESUMO

As redes de sensores sem fio possuem recursos bastante restritos, tais como largura de banda e capacidade computacional limitadas, além das restrições de memória e, principalmente, de energia. Devido a essas restrições, enviar grandes quantidades de dados pode ser problemático, causando atraso demasiado no tempo de resposta e também degradando mais rapidamente o tempo de vida da rede, pois o consumo de energia é maior. Os dados que trafegam nessas redes podem ser univariados, que são representados por um único conjunto de valores de um mesmo tipo de sensor, ou multivariados, que são representados por conjuntos de valores de um ou mais nós. Devido às restrições das redes de sensores sem fio, uma estratégia que tem recebido crescente atenção por parte de pesquisadores é a redução de dados, que visa diminuir a quantidade de dados que são enviados para o sorvedouro. A redução pode ser aplicada no momento do sensoriamento ou através de um nó líder e diferentes técnicas podem ser utilizadas para esse fim. Entre as principais técnicas empregadas para efetuar redução de dados em redes de sensores sem fio é possível destacar agregação de dados, amostras adaptativas, técnicas de *stream* de dados e redução multivariada. Nessa direção, este trabalho propõe um algoritmo de amostragem para redução de dados multivariados em redes de sensores sem fio. É apresentada uma solução geral, utilizando técnicas de análise de componentes, para fazer uma classificação ordenada dos dados sensoriados, possibilitando a seleção de uma amostra contendo apenas os dados mais relevantes para a aplicação. A classificação dos dados é feita com base nos escores da primeira componente obtida pela técnica de análise de componentes e para a ordenação pode considerar os valores superiores, inferiores ou intermediários desses escores, de acordo com a aplicação. O algoritmo é avaliado em função da representatividade dos dados e do comportamento da rede, sendo simulado no programa R e no *Network Simulator 2*, respectivamente. As avaliações são feitas considerando dois cenários. No primeiro, considera-se um nó que possui diferentes sensores monitorando fenômenos distintos e no segundo, um nó que processa as informações de diferentes nós da rede que monitoram o mesmo fenômeno. No caso da representatividade dos dados, foram gerados conjuntos de dados sintéticos e pseudo-reais, utilizando as distribuições *normal* e *skew-normal* multivariadas, com e sem a presença de ruído nos dados, e aplicando-se os testes *ANOVA* - **AN***alysis* **O***f* **VA***riance* - e erro absoluto relativo para avaliar o desempenho do algoritmo. Resultados mostram que o algoritmo, utilizando ambas as técnicas avaliadas, tem desempenho satisfatório em praticamente todos os casos, sendo que em muitos deles, os erros observados foram próximos de zero. No caso do comportamento da rede, os resultados mostram que o

uso da técnica proposta reduz consideravelmente o consumo de energia e o atraso no envio dos pacotes na rede.

Palavras-chave: Redes de sensores sem fio. Redução de dados multivariados. Análise de componentes principais. Análise de componentes independentes.

ABSTRACT

In wireless sensor networks there are many restrictions, such as limited bandwidth and computational capability, memory and mainly energy restrictions. Due these restrictions, to send high amount of data can be problematic, causing excessive delay in response time and diminishing the network lifetime because the energy consumption is higher. The data which traffic in these networks can be univariate, that are represented by a single set of values of a same kind of sensor, or multivariate, that are represented by sets of values of one or more nodes. Due to the restrictions of the wireless sensor networks, a strategy which has received increasing attention from researchers is the data reduction, which aims to reduce the amount of data that are sent to the sink. The reduction can be applied at the sensing or through a leader node and different techniques may be used for this purpose. Among the main techniques used to perform reduction of data in wireless sensor networks can highlight data aggregation, adaptive sampling, data stream techniques and multivariate reduction. In this way, this paper proposes a sampling algorithm for multivariate data reduction in wireless sensor networks. It is presented a general solution, using component analysis techniques, that create a rank of the sensory data, allowing the selection of a sample containing only the data most relevant to the application. The data rank is based on the scores of the first component obtained by the component analysis technique and for the ordering can to consider the upper, lower or intermediate scores, according to the application. The algorithm is evaluated in terms of the data representativeness and of the network behavior, being simulated in the R program and in the Network Simulator 2, respectively. Evaluations are made in two scenarios. At first, it is considered a node that has different sensors monitoring different phenomena and the second, a node that processes the information from different nodes that monitor the same phenomenon.In the case of the data representativeness, were generated synthetic and pseudo-real data sets, using the normal and skew-normal multivariate distributions, with and without the presence of noise in the data, and applying the ANOVA - **AN**alysis **O**f **VA**riance - test and absolute relative error to evaluate the performance of the algorithm. Results show that the algorithm, using both evaluated techniques, has satisfactory performance in almost all cases, being many of the errors observed close to zero. Regards the network behavior, results show that using the proposed technique, the energy consumption and delay are diminished considerably.

Keywords: Wireless sensor networks. Multivariate data reduction. Principal component analysis. Independent component analysis.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

1.1 Motivação

O mundo ao nosso redor possui uma variedade de fenômenos que podem ser descritos por algumas grandezas, tais como temperatura, pressão e umidade, que podem ser monitorados por dispositivos com poder de sensoriamento, processamento e comunicação (AQUINO, 2008). Tais dispositivos, trabalhando de forma cooperativa, constituem as redes de sensores sem fio (AKYILDIZ et al., 2002; ARAMPATZIS; LYGEROS; MANESIS, 2005; ESTRIN et al., 1999). As redes de sensores sem fio são um tipo especial de redes *adhoc* (ROYER; TOH, 1999), compostas por um conjunto de sensores, que realizam processamento local e sensoriamento. Essas redes podem ser compostas ainda por nós atuadores, que interferem no ambiente monitorado e que podem ser fisicamente os próprios nós sensores. Outros componentes das redes de sensores sem fio são os sorvedouros, que são dispositivos que apresentam capacidade computacional superior a dos demais nós e têm a responsabilidade de receber e processar as informações de sensoriamento, e os *gateways*, responsáveis por prover a comunicação com outras redes. A estrutura de uma rede de sensores sem fio é ilustrada na figura 1.

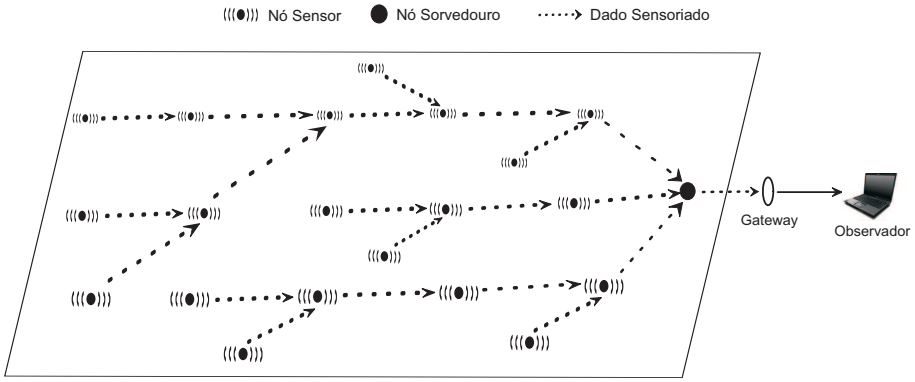


FIGURA 1: Estrutura de uma rede de sensores sem fio

Uma característica que distingue estas redes das demais é que os nós sensores possuem recursos bastante restritos, tais como largura de banda e capacidade computacional limitadas, além das restrições de memória e, principalmente, de energia. Os sensores que compõem os nós da rede são equipados com bateria e, em muitas aplicações, eles serão colocados em áreas remotas, impossibilitando o acesso a esses elementos para manutenção. Neste cenário, o tempo de vida da rede depende da quantidade de energia disponível nos nós sensores e, por

isso, esses nós devem balancear seus recursos limitados com o objetivo de aumentar o tempo de vida da rede (MINI; LOUREIRO, 2009).

Devido às restrições mencionadas, enviar grandes quantidades de dados pode ser problemático, causando atraso demasiado no tempo de resposta e também degradando mais rapidamente o tempo de vida da rede, pois o consumo de energia é maior e a energia armazenada nas baterias dos sensores esgota-se com maior rapidez. Com isso, uma abordagem que vem recebendo muita atenção por parte dos pesquisadores é a redução de dados, que possui um conjunto de técnicas para diminuir a quantidade de informações transmitidas até o sorvedouro. Dentre as técnicas utilizadas para efetuar redução de dados em redes de sensores sem fio é possível destacar:

- **Agregação de dados:** a utilização de agregação como técnica de redução visa combinar os dados vindos de diferentes fontes, eliminando redundância, minimizando o número de transmissões e economizando energia (KRISHNAMACHARI; ESTRIN; WICKER, 2002; DASGUPTA; KALPAKIS; NAMJOSHI, 2003).

- **Amostragem:** essa técnica visa selecionar um subconjunto do conjunto de dados originais. Uma técnica de amostragem usual é denominada amostragem adaptativa, que modifica a forma de sensoriamento, de acordo com os requisitos da aplicação e os recursos dos nós, com o objetivo de propagar apenas uma amostra contendo a informação mais relevante para a aplicação (MARBINI; SACKS, 2003; SANTINI; ROMER, 2006).

- **Técnicas baseadas em stream de dados:** em redes de sensores sem fio, alguns fenômenos monitorados geram dados com características que possibilitam classificá-los como *stream* de dados. O *stream*, nesse caso, possui tamanho moderado, é impreciso e com ruído e representa as amostras da população de interesse (AQUINO, 2008; AQUINO et al., 2008).

Além dessas técnicas, é possível citar ainda a fusão de dados (NAKAMURA et al., 2005; NAKAMURA; LOUREIRO; FRERY, 2007), a compressão de dados (KIMURA; LATIFI, 2005; PATTEM; KRISHNAMACHARI; GOVINDAN, 2008) e o processamento colaborativo (YU; PRASANNA, 2005; WANG; WANG, 2007). Neste trabalho, a redução de dados é efetuada através da técnica de amostragem.

Uma consideração importante sobre os dados que trafegam nas redes de sensores sem fio diz respeito a sua classificação. Com base na quantidade de fenômenos monitorados, os dados sensoriados podem ser classificados como univariados ou multivariados. Dados univariados são

representados por um único conjunto de valores de um mesmo tipo de sensor, por exemplo, um nó que monitora apenas a temperatura do ambiente. Por outro lado, os dados multivariados são representados por conjuntos de valores de um ou mais nós, por exemplo, um nó que possui sensores que monitoram temperatura, pressão e umidade simultaneamente, ou um nó que processa os dados de um conjunto de nós que monitoram apenas temperatura (AQUINO, 2008).

Um exemplo de aplicação na qual os dados são multivariados é o monitoramento da qualidade do ar. Considerando esse tipo de aplicação, algumas cidades brasileiras, como Rio de Janeiro e São Paulo, possuem estações de monitoramento nas quais sensores são utilizados para obter informações a respeito da qualidade do ar (ALBUQUERQUE, 2007). Porém, essas estações não são caracterizadas como redes de sensores sem fio, pois não possuem comunicação entre os nós nem processamento local. Nessas aplicações, a qualidade do ar é monitorada em inúmeras áreas, onde são medidos diferentes parâmetros, como óxido de nitrogênio, dióxido de nitrogênio, entre outros. Como diferentes variáveis são monitoradas, nesse caso, é possível caracterizar os dados como multivariados.

Um ponto importante a respeito das aplicações em redes de sensores que geram dados multivariados é que ainda poucos trabalhos têm sido desenvolvidos considerando esse tipo de dados. Na redução de dados multivariados são utilizados métodos para estimar o comportamento do dado multivariado a ser sensoriado e apenas as diferenças observadas ao longo do tempo ou pequenas amostras são enviadas (SEO; KANG; RYU, 2005; JUNIOR et al., 2009). No caso do monitoramento de diferentes fenômenos, a partir da determinação da correlação entre esses fenômenos, torna-se possível a utilização de técnicas estatísticas para classificar os dados sensoriados de acordo com sua correlação e/ou variância. Dessa forma, é possível realizar um processo de amostragem, reduzindo a quantidade de dados enviados para processamento.

1.2 Objetivo

O problema de redução multivariada em redes de sensores sem fio pode ser enunciado como segue: *"Como efetuar a redução de um conjunto de dados multivariados em redes de sensores sem fio, preservando suas características e correlação?"*

Dentre as possíveis técnicas estatísticas para classificar de forma ordenada os dados multivariados e possibilitar o processo de redução baseada em amostragem, é possível destacar a análise de componentes. Dessa forma, duas hipóteses são consideradas nesse caso. Primeiramente, a utilização de técnicas de análise de componentes para classificar os dados pode ser

utilizada para auxiliar na redução dos dados, mantendo a sua representatividade. A segunda hipótese é que através da redução é possível diminuir o consumo de energia e o atraso no envio das mensagens na rede. Diferentes técnicas de análise de componentes podem ser empregadas nesse caso, tais como:

- **Análise de componentes principais:** realiza uma transformação linear em um conjunto de dados, com o objetivo de obter um novo conjunto de dados com dimensão reduzida (PEARSON, 1901).
- **Análise de componentes principais robusta:** técnica desenvolvida com o intuito de tratar possíveis deficiências da técnica de análise de componentes principais tradicional no que se refere à presença de valores discrepantes ou atípicos nos dados (XU; YUILLE, 1995).
- **Análise de componentes independentes:** também realiza uma transformação linear nos dados, considerando que suas componentes são estatisticamente independentes. Pode reduzir, aumentar ou manter o número de dimensões dos dados originais (COMON, 1994).

Este trabalho propõe um algoritmo de redução baseado em amostragem para dados multivariados em redes de sensores sem fio. Neste trabalho, diferentemente da maioria dos trabalhos que utilizam técnicas de análise de componentes, as mesmas não são utilizadas para reduzir a dimensão dos dados, mas para classificá-los de forma ordenada, considerando os escores da primeira componente, possibilitando uma amostragem contendo os dados mais relevantes dessa classificação. O uso das técnicas de análise de componentes dessa forma, auxilia na redução dos dados mantendo sua representatividade, pois é possível eliminar redundâncias e detalhes pouco significativos, ou seja, é possível eliminar dependências lineares e, consequentemente, dados que não são relevantes para a aplicação. É importante destacar que o objetivo principal deste trabalho é apresentar uma solução para o problema de redução de dados multivariados em redes de sensores sem fio, tendo como métrica as avaliações da representatividade dos dados reduzidos, do consumo de energia e do atraso na rede. Dependendo da aplicação a ser considerada, pode ser necessário avaliar outras métricas e cenários não contemplados neste trabalho.

O objetivo é propor uma solução geral para redução de dados multivariados em redes de sensores sem fio baseada em técnicas de análise de componentes. Essa proposta pode ser utilizada para diferentes aplicações que geram dados multivariados, sejam eles correlacionados ou não. O algoritmo proposto, além de poder utilizar diferentes técnicas baseadas em análise

de componentes, pode ter parâmetros facilmente ajustados para obter melhor desempenho no que se refere à qualidade dos dados reduzidos. Por exemplo, é possível modificar a forma de selecionar os dados mais relevantes para a aplicação, com a utilização dos escores superiores, inferiores ou intermediários da primeira componente.

1.3 Formulação do problema

Considerando o que foi discutido na seção anterior, o problema de redução de dados multivariados em redes de sensores sem fio pode ser modelado de acordo com o diagrama mostrado na figura 2 (AQUINO et al., 2008). Nesse diagrama, $\mathcal{N}$ representa o ambiente e o processo a ser medido, F é o fenômeno multivariado de interesse e $\mathcal{V}^*$, seu domínio espaço-temporal. Se uma observação for completada sem problemas, ou seja, sem possibilidades de perdas de informações, tem-se um conjunto de regras ($\mathcal{R}^*$) ideais para tomada de um conjunto de decisões ideais (D^*).

$$\begin{array}{ccccccc} \mathcal{N} & \xrightarrow{F} & \mathcal{V}^* & \xrightarrow{S} & \mathcal{V} & \xrightarrow{\Psi} & \mathcal{V}' \\ & & \downarrow{\scriptstyle \mathcal{R}^*} & & \downarrow{\scriptstyle \mathcal{R}} & & \downarrow{\scriptstyle \mathcal{R}'} \\ & & D^* & & D & & D' \end{array}$$

FIGURA 2: Modelagem do problema de redução de dados multivariados em redes de sensores sem fio

Contudo, em redes de sensores sem fio, no caso do monitoramento de dados multivariados, ao invés de uma situação ideal, tem-se um conjunto de nós sensores, $S = (\{S^1_{1..n}, \ldots, S^s_{1..n}\})$, monitorando s fenômenos e produzindo conjuntos de amostras no domínio $\mathcal{V}_i$, com $1 \leq i \leq s$, sendo o conjunto de dados multivariados denotado por $\mathcal{V} = (\{\mathcal{V}^1_{1..n}, \ldots, \mathcal{V}^s_{1..n}\})$. Usando todos os dados coletados pelos sensores, é possível conceber um conjunto de regras ($\mathcal{R}$) para prover um conjunto de decisões (D). Devido à complexidade dos sistemas nos quais as redes de sensores atuam, é impraticável utilizar um conjunto de regras gerais $\mathcal{R}$ que possam ser aplicadas em qualquer sistema. Tanto as regras, como as decisões a serem tomadas, são específicas para cada aplicação.

Conforme discutido anteriormente, enviar grandes quantidades de dados pode ser muito oneroso em termos de energia, largura de banda e tempo de entrega das mensagens, causando atraso demasiado e degradando mais rapidamente o tempo de vida da rede. Dessa forma, utilizar todo o conjunto $\mathcal{V}$ pode ser inviável e a utilização de técnicas de redução de dados

pode auxiliar na solução desse problema. Na figura 2, as técnicas de redução de dados multivariados nos sensores são representadas por

$$\Psi : \mathbb{R}^{s \times n} \rightarrow \mathbb{R}^{s \times n'} \mid n' < n, \tag{1.1}$$

onde n representa o número de dados coletados por cada sensor s, n' é o número de dados do conjunto reduzido e o cálculo da redução Ψ é dado por

$$\Psi = \psi_A \circ \psi_O \circ \psi_C, \tag{1.2}$$

onde ψ_C representa o cálculo das componentes, ψ_O é a ordenação dos escores da primeira componente e ψ_A representa a amostragem. O conjunto de dados reduzidos do domínio $\mathcal{V}$ é representado por $\mathcal{V}'^{s}_{n'}$. Com isso, as novas regras que usam $\mathcal{V}'$ são representadas por $\mathcal{R}'$, e as mesmas conduzem a um conjunto de decisões D'. Considerando a necessidade de se efetuar uma redução Ψ no conjunto $\mathcal{V}$, um ponto fundamental a ser analisado é se a partir do conjunto reduzido $\mathcal{V}'$ e de acordo com um conjunto de regras $\mathcal{R}'$ é possível tomar um conjunto de decisões D', sendo D' equivalente ao conjunto de decisões D, que seriam tomadas caso fosse utilizado todo o conjunto de dados $\mathcal{V}$. Nesse caso, a redução Ψ recebe como entrada uma matriz de dados $\mathcal{V}^{s}_{n}$ e retorna uma saída $\mathcal{V}'^{s}_{n'}$, onde $n' < n$, como definido anteriormente. É importante destacar que as decisões dependem da aplicação. Por exemplo, em uma aplicação de monitoramento de uma floresta, a partir de determinado valor de temperatura, conclui-se que está havendo um incêndio. Neste trabalho o foco está nas regras, que serão a base para a determinação da equivalência entre D e D', uma vez que não há comparações com nenhum sistema de tomada de decisões.

1.4 Organização do trabalho

Este trabalho é organizado como segue. No capítulo 2, são apresentados os fundamentos teóricos importantes para uma melhor compreensão da abordagem proposta neste trabalho. Os trabalhos relacionados às principais técnicas de redução de dados em redes de sensores sem fio são descritos no capítulo 3. No capítulo 4, são apresentados a proposta para redução de dados multivariados através de amostragem baseada em análise de componentes e o algoritmo de amostragem proposto. No capítulo 5, discorre-se a respeito dos resultados de simulações referentes à análise da representatividade dos dados reduzidos em relação aos originais, utilizando a distribuição *normal*. A avaliação do comportamento da rede com a utilização do algoritmo de amostragem proposto é apresentada no capítulo 6. Por fim, no capítulo 7, são

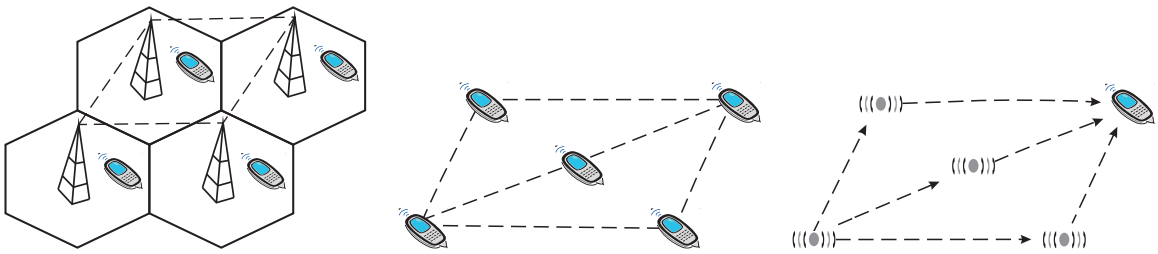
apresentadas as conclusões do trabalho e futuras direções e, no Anexo, são apresentados os resultados da avaliação da representatividade dos dados utilizando a distribuição *skew-normal*.

2 FUNDAMENTOS

Este capítulo tem como objetivo apresentar os principais fundamentos necessários para uma melhor compreensão deste trabalho. Para isso, serão discutidos conceitos relacionados às redes de sensores sem fio, às técnicas para processamento de dados multivariados baseadas em análise de componentes e ainda às técnicas utilizadas para geração e avaliação de dados multivariados.

2.1 Redes de sensores sem fio

Além das características mencionadas no capítulo 1, outros conceitos são importantes para um melhor entendimento das redes de sensores sem fio, e os mesmos são apresentados nesta seção. Para entender as redes de sensores sem fio é necessário considerar as redes estruturadas e *ad-hoc*. Nas redes estruturadas, ilustrada na figura 3(a), existe uma estação base que é responsável pela comunicação entre os nós da rede. Por outro lado, nas redes *ad-hoc* (figura 3(b)), a comunicação é feita através dos nós que se encontram entre a origem e o destino, não necessitando, portanto, da estação base para prover essa comunicação. Assim como as redes *ad-hoc*, as redes de sensores sem fio (figura 3(c)) não necessitam de uma estação base para comunicação entre dois nós. O objetivo dessas redes é processar e propagar os dados sensoriados a um observador externo à rede. As redes de sensores são muitas vezes empregadas em ambientes hostis e com condições imprevisíveis, devendo, portanto, ser auto-configuráveis, adaptáveis e com gerenciamento escalável (AKYILDIZ et al., 2002).



(a) Rede estruturada (b) Rede ad-hoc (c) Rede de sensores

FIGURA 3: Tipos de redes sem fio

Os nós sensores possuem tamanhos reduzidos e, com isso, uma arquitetura simples e com limitações de processamento e armazenamento. Esses dispositivos são formados por

quatro componentes básicos: uma unidade perceptiva, que pode possuir alguns sensores e um conversor de sinais analógicos para digitais (ADC); uma unidade de processamento, com memória e processador; um transceptor; e uma fonte de energia, que normalmente não é renovável. Além disso, podem existir ainda elementos que complementam essa estrutura, como sistema de localização, mecanismo de mobilidade e gerador de energia. Essa estrutura básica dos nós sensores é apresentada na figura 4 (AKYILDIZ et al., 2002).

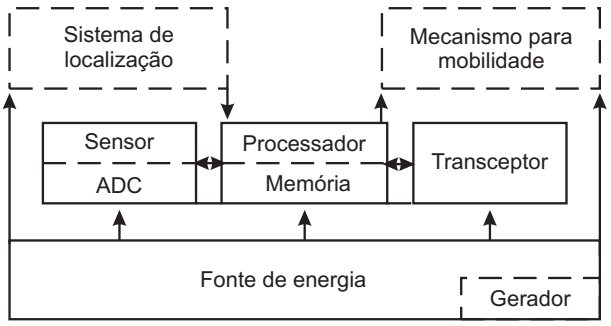


FIGURA 4: Estrutura básica do nó sensor

De acordo com seus elementos básicos, com a disposição dos nós na área de sensoriamento e com a maneira de monitorar os fenômenos, as redes de sensores sem fio podem receber diferentes classificações (TILAK; ABU-GHAZALEH; HEINZELMAN, 2002). Caso a rede possua agrupamentos de nós, existindo um nó líder para representar esses agrupamentos, a rede é denominada hierárquica, caso contrário, ela é denominada plana. Uma rede é homogênea, caso os nós possuam a mesma configuração de *hardware*, ou heterogênea, caso os nós possuam configurações diferentes. Quando os nós da rede possuem o mesmo raio de comunicação, a rede é chamada simétrica, caso contrário, ela é assimétrica. Além disso, uma rede pode ser contínua, caso os dados sensoriados sejam enviados continuamente, ou programada, quando existe uma programação pré-estabelecida para envio dos dados. Por fim uma rede de sensores sem fio pode ser dirigida a eventos, caso o envio de dados seja efetivado somente na ocorrência de algum evento, ou sob demanda, quando é permitida a consulta total ou parcial dos dados a qualquer momento. Neste trabalho, são utilizadas tanto as redes planas quanto as hierárquicas, sendo elas homogêneas, simétricas e contínuas.

As redes de sensores sem fio possuem cinco funcionalidades básicas (LOUREIRO et al., 2003). A primeira funcionalidade consiste na configuração inicial da rede e é denominada estabelecimento. Outra funcionalidade, responsável pela adaptação da rede às mudanças de configurações que ocorrem com o passar do tempo, é chamada manutenção. O sensoriamento é a funcionalidade responsável pela coleta de dados no ambiente. Existem ainda o processamento dos dados a serem transmitidos para o sorvedouro e a comunicação, que é a funcionalidade encarregada do envio desses dados. Os algoritmos considerados neste trabalho são executados

na fase de processamento.

De forma geral, é possível destacar dois tipos de aplicações em redes de sensores sem fio. Aplicações de monitoramento, cujo objetivo é obter característica(s) do ambiente sem interferir no mesmo, onde os nós apenas processam os dados. Nessas aplicações, as redes de sensores sem fio podem ser empregadas para diversos fins, como monitoramento ambiental, climático, de tráfego de veículos, entre outros (ARAMPATZIS; LYGEROS; MANESIS, 2005). Destacam-se ainda as aplicações de atuação, que podem interferir no ambiente monitorado. Nesse caso, os nós assumem o papel de atuador, recebendo requisições de um elemento externo para interferir em tal ambiente. Um exemplo é a utilização de redes de sensores sem fio na agricultura, por exemplo, onde os sistemas de irrigação são acionados de acordo com a umidade do solo (MCCULLOCH et al., 2008).

Em consequência das diversas áreas em que têm sido empregadas, as redes de sensores sem fio vêm recebendo cada vez mais atenção de pesquisadores com o intuito de estudar seus principais problemas. As redes de sensores sem fio diferem-se das redes *ad-hoc* por possuírem, potencialmente, centenas de milhares de nós (SOHRABI et al., 2000). Os nós sensores podem operar em ambientes inóspitos e para alcançar uma boa resolução de sensoriamento (ou seja, uma área de cobertura adequada), essas redes devem ser densas, o que torna a escalabilidade em relação ao número de nós um fator essencial. A densidade depende da aplicação a ser considerada, sendo que um exemplo que pode ser considerado é a existência de 21 vizinhos por nó (VILLAS et al., 2009). Esse número pode ser considerado, por exemplo, para aplicações de monitoramento ambiental.

Diferentemente das redes *ad-hoc* convencionais, na maioria das vezes, os nós de uma rede de sensores sem fio permanecem estacionários (imóveis) após sua deposição (colocação na área a ser monitorada), não impedindo, entretanto, que essas redes possuam topologia dinâmica. Durante períodos de baixa atividade, muitos nós entram em estado de dormência, ou seja, se desligam para economizar energia. Essa operação de ligar e desligar os rádios efetuada pelos nós sensores é um dos fatores que tornam a topologia dinâmica. Além disso, os nós podem ficar fora de serviço devido ao esgotamento da energia de suas baterias ou ainda por causa de um evento destrutivo. Existe também a possibilidade de utilização de nós móveis, como ocorre, por exemplo, em aplicações de monitoramento subaquático (POMPILI; MELODIA; AKYILDIZ, 2006). Considera-se neste trabalho que os nós sempre permanecem estacionários após sua deposição.

2.2 Análise de componentes

Nessa seção, discorre-se a respeito de algumas técnicas baseadas em análise de componentes, apresentando os principais fundamentos necessários para um melhor entendimento das mesmas. Essa descrição é necessária, uma vez que o algoritmo de amostragem proposto neste trabalho pode utilizar diferentes técnicas de análise de componentes para efetuar a tarefa de classificar os dados sensoriados, auxiliando no processo de obtenção dos dados mais relevantes para a aplicação.

2.2.1 Componentes principais

Análise de Componentes Principais – *Principal Component Analysis* – (PCA) (PEARSON, 1901; HOTELLING, 1933; JACKSON, 2003), também conhecida como transformação de Karhunen-Loève, é uma das ferramentas mais poderosas para o tratamento de dados multivariados. É uma transformação entre espaços γ-dimensionais, derivada da matriz de covariância dos dados de entrada gerando um novo conjunto de dados, de modo que cada valor resultante é uma combinação linear dos valores originais. Essas combinações lineares são denominadas componentes principais. O número de componentes principais é igual ao número de dimensões dos dados originais e esses podem ser ordenados de acordo com a sua variância. Com isso, a primeira componente principal explica a maior parte da variância total da matriz de entrada, ou seja, é aquela que provê mais informações. Por outro lado, a última componente principal tem a menor variância, ou seja, menor quantidade de informações.

Após a determinação das componentes principais, os valores numéricos ou escores de cada uma dessas componentes podem ser calculados e posteriormente analisados, através de técnicas estatísticas tais como análise de variância, análise de regressão, entre outras. Em geral, a utilização de PCA se dá para reduzir a quantidade de dimensões dos dados a serem avaliadas. Dessa forma, entre as s componentes principais calculadas, são escolhidas k componentes, onde $k < s$. Essas k componentes são escolhidas de forma a explicarem a maior parte da variância global dos dados originais, permitindo uma análise sem perdas significativas de qualidade. Além disso, segundo Mingoti (2005):

> É comum utilizar os escores das componentes principais para condução de análise estatística de dados ou para a simples ordenação (*ranking*) dos elementos amostrais observados, com o intuito de identificar aqueles que estão com maiores, ou menores, valores globais das componentes.

A propriedade mais importante do novo conjunto de dados gerado por PCA, ou seja, das componentes principais, é que essas componentes são não-correlacionadas entre si (JACKSON,

2003), garantindo dessa forma que não haja redundância entre os dados e que seja obtido um novo conjunto de dados com propriedades para análise multivariada. Dessa forma, a técnica consegue reduzir a dimensão dos dados mantendo a qualidade dos mesmos. A transformação de componentes principais pode ser descrita nas seguintes etapas:

1. Calcular Σ, a matriz de covariância dos dados (supõe-se que ela seja definida positiva pois trata-se de variâncias).
2. Decompor Σ nos autovetores U e autovalores λ. Essa matriz será diagonalizável, uma vez que a matriz de covariância é definida positiva (KRZANOWSKI, 1995).
3. Calcular o novo conjunto de dados, multiplicando o valor de cada variável pela matriz dos autovetores.

Os autovalores representam o comprimento dos eixos das componentes principais do conjunto de dados e são medidos na unidade da variância. Associado a cada autovalor, existe um vetor de módulo unitário chamado autovetor. Os elementos de cada autovetor são fatores de ponderação que definem a contribuição da variável da matriz de dados original para uma componente principal, numa combinação linear. Os autovetores representam as direções dos eixos das componentes principais.

Como descrito na formulação do problema, na seção 1.3, $\mathcal{V}$ representa um conjunto de dados sensoriados. Assim, o método de componentes principais pode ser formulado da seguinte forma: dada uma matriz de dados originais $\mathcal{V}$, com s variáveis correlacionadas, aplicar PCA consiste em calcular a matriz C, que possui s variáveis não correlacionadas, de forma que cada componente principal será calculada por

$$C_i = u_i'[\mathcal{V} - \overline{\mathcal{V}}], \tag{2.1}$$

onde para cada $1 \leq i \leq s$, $u_i = (u_{i,1}, \ldots, u_{i,s})$ é o autovetor i da matriz de covariância dos dados $\mathcal{V}$.

Outra propriedade importante do PCA é que a equação (2.1) pode ser invertida, restaurando as variáveis originais em função das componentes principais. Para isso utiliza-se

$$\mathcal{V} = \overline{\mathcal{V}} + UC. \tag{2.2}$$

Devido ao autovetor U ser ortonormal, temos $U^{-1} = U'$; com isto, dada a matriz C, os dados originais $\mathcal{V}$ podem ser unicamente determinados pela equação (2.2). É importante ressaltar que neste trabalho não é necessário fazer essa reconstrução dos dados originais, uma vez que

não é feita redução de dimensionalidade dos dados, mas apenas a utilização de C para efetuar a classificação.

PCA pode apresentar deficiências no que se refere à robustez do método quando da presença de valores discrepantes ou atípicos nos dados de entrada. Para tentar resolver esse problema de robustez do método, existe a técnica PCA-robusta (XU; YUILLE, 1995; SONG; SHAOWEI, 1997; TORRE; BLACK, 2001). A principal diferença dessa técnica para a técnica PCA tradicional está na maneira de se calcular a matriz de covariância Σ dos dados, que no caso de PCA-robusta pode ser feita calculando-se cada elemento de Σ separadamente, através de estimadores robustos específicos para o coeficiente de correlação ou covariância, ou ainda, estimando Σ como um todo.

Análise de Componentes Principais tem sido utilizada em inúmeras áreas como a física, matemática, geografia e, principalmente para o tratamento de imagens. Em redes de sensores sem fio, a técnica pode ser utilizada para classificar de forma ordenada os dados multivariados sensoriados com base nos escores da primeira componente principal e, com essa classificação, selecionar uma pequena amostra dos dados originais para se transmitir até o sorvedouro (JUNIOR et al., 2009). O uso de PCA, nesse caso, pode auxiliar na redução dos dados mantendo sua representatividade, pois é possível eliminar redundâncias e dados pouco relevantes para a aplicação. Assim como no caso da utilização de PCA tradicional, os escores da primeira componente principal obtida pela técnica PCA-robusta serão utilizados para classificar os dados sensoriados, para possibilitar ao algoritmo de amostragem proposto neste trabalho selecionar uma amostra contendo os dados mais relevantes para a aplicação.

2.2.2 Componentes independentes

Outra técnica utilizada para o tratamento de dados multivariados é denominada Análise de Componentes Independentes (*Independent Component Analysis* - ICA) (COMON, 1994; HYVÄRINEN, 1999; HYVäRINEN; KARHUNEN; OJA, 2001). O objetivo dessa técnica é encontrar uma transformação na qual as componentes C_i são estatisticamente independentes umas das outras. De acordo com (HYVÄRINEN, 1999), ICA pode ser considerada uma técnica para reduzir redundância. A definição mais simples e mais utilizada por pesquisadores considera que a transformação linear através de ICA consiste em estimar o seguinte modelo geral para os dados:

$$\mathcal{V} = AC, \tag{2.3}$$

onde $\mathcal{V}$ são os dados de entrada, A é uma "matriz de mistura" retangular e C_i no vetor $C = (C_1, ..., C_n)^T$, assim como no caso de PCA representam as componentes, que nesse

caso, diferentemente de PCA, são consideradas independentes. Assume-se ainda que C possui média zero e covariância finita.

Algumas restrições devem ser consideradas para se assegurar a identidade do modelo ICA (HYVÄRINEN, 1999). A primeira restrição considera que apenas uma das componentes independentes C_i pode ser gaussiana, todas as outras devem ser não gaussianas. Isso porque, para variáveis aleatórias gaussianas, a simples inexistência de correlação implica em independência, e dessa forma, qualquer representação não-correlacionada resultaria em componentes independentes. Contudo, se mais de uma das componentes são gaussianas, ainda existe a possibilidade de se identificar as componentes independentes não gaussianas, bem como as correspondentes colunas da "matriz de misturas". Outra restrição é que todas as colunas da matriz A devem ser linearmente independentes.

Componentes independentes versus componentes principais

ICA possui algumas semelhanças e diferenças em relação à PCA. Assim como PCA, ICA tem o objetivo de realizar uma transformação linear em um conjunto de dados. As diferenças começam pelas considerações contraditórias em relação à gaussianidade, uma vez que PCA assume que dados são gaussianos, enquanto ICA assume que são não gaussianos. Além disso, a aplicação principal de PCA é para a redução de dimensionalidade, enquanto ICA pode reduzir, aumentar ou manter o número de dimensões constante.

Conforme descrito em (HYVÄRINEN, 1999), diferentemente de PCA, a definição de ICA não implica em uma ordenação das componentes independentes em relação à variância dos dados. Embora seja possível determinar a ordem das componentes, por exemplo, de acordo com a não gaussianidade das componentes, usualmente considera-se que as componentes têm a mesma variância. Dessa forma, neste trabalho assume-se que a primeira componente independente possui a maior variância.

Outro ponto importante da técnica ICA é que, de forma semelhante à PCA, em cada componente independente, é possível determinar os escores dessas componentes e, a partir desses escores, realizar análises estatísticas dos dados. Dessa forma, assim como na utilização de PCA e PCA-robusta, os escores da primeira componente independente serão utilizados para efetuar uma classificação ordenada dos dados coletados pelos sensores, para que o algoritmo de amostragem proposto neste trabalho selecione uma amostra contendo os dados mais relevantes para a aplicação. Para a determinação desses escores das componentes independentes será utilizado o algoritmo FastICA (HYVäRINEN; OJA, 1997; HYVäRINEN, 1999).

2.3 Geração e análise da representatividade dos dados

Utilizou-se para a geração dos dados $\mathcal{V}_n^s$ as distribuições *normal* e *skew-normal* multivariadas. A distribuição *normal* foi utilizada para simular as aplicações nas quais os dados são simétricos, ou seja, que possuem um comportamento padrão de variação. Entretanto, em algumas aplicações os dados não são simétricos e, portanto, a utilização apenas da distribuição *normal* não seria adequada para a simulação. Dessa forma, foi utilizada também a distribuição *skew-normal* (AZZALINI; VALLE, 1996; AZZALINI; CAPITANIO, 1999), que efetua a geração de dados assimétricos, garantindo assim, uma fidelidade maior com as aplicações consideradas.

Esses dados são gerados através do programa estatístico R[†], com e sem a presença de ruído nos mesmos. No caso da geração de dados com ruído, eles são primeiramente gerados utilizando as distribuições *normal* e *skew-normal* e, posteriormente, é introduzido um ruído aleatório de até 400% nos dados gerados. Além da geração dos dados nas distribuições mencionadas e da introdução de ruído nos dados, os cálculos das componentes principais e independentes também são feitos a partir de algoritmos disponíveis no programa R. Após a geração dos dados e execução do algoritmo de redução, são utilizados o teste de hipótese *Analysis of Variance - ANOVA* (THOMSON, 1993) e o erro absoluto relativo (FRERY et al., 2008), para determinar a qualidade dos dados reduzidos. Nesse caso, o teste *ANOVA* e o erro absoluto relativo representam as regras $\mathcal{R}'$, utilizadas para se tomar o conjunto de decisões D', como definido na seção 1.3. Esses testes são efetuados conforme descrito a seguir:

- **Teste de hipótese**: utilizou-se o teste de hipótese *ANOVA*. Esse teste tem por objetivo avaliar se existem diferenças estatisticamente significativas entre as médias do conjunto de dados original e do conjunto de dados reduzido. O cálculo é dado por,

$$F = D_B^2 / D_W^2, \tag{2.4}$$

  onde D_B^2 representa a dispersão entre os conjuntos $\mathcal{V}$ e $\mathcal{V}'$ e D_W^2 a dispersão dentro dos conjuntos. A partir desse cálculo, o $p - valor$ é utilizado para decidir se a hipótese nula H_0 deve ser aceita ou rejeitada. Nesse caso, aceitar a hipótese nula indica que não existem diferenças significativas entre as variâncias dos dois conjuntos. Resultados para o $p - valor$ acima de 0,05 são considerados satisfatórios para aceitação da hipótese nula. Por convenção será utilizado $\mathcal{R}'_\Phi$ para indicar a utilização desse teste como regra de decisão.

- **Erro absoluto relativo**: o erro absoluto relativo considera a comparação entre as

[†] *The R Project for Statistical Computing*. Disponível em: http://www.r-project.org/

médias dos dados originais $\mathcal{V}$ e reduzidos $\mathcal{V}'$. Esse erro é dado por,

$$\mathcal{R}'_{\Upsilon} = 100\, Max\{\forall_i \left|(\overline{\mathcal{V}_i} - \overline{\mathcal{V}'_i})/\overline{\mathcal{V}_i}\right|\}. \tag{2.5}$$

Note que $\mathcal{R}'_{\Upsilon}$ é calculado para cada sensor i e somente o maior deles, situação onde a técnica foi menos eficiente, será utilizado.

2.4 Conclusões parciais

Este capítulo apresentou os principais fundamentos necessários para um melhor entendimento da solução proposta neste trabalho. Estes fundamentos foram apresentados levando em consideração a formulação do problema de redução de dados multivariados em redes de sensores sem fio descrita no capítulo anterior. Nos próximos capítulos serão apresentados alguns trabalhos relacionados à redução de dados em redes de sensores sem fio, a solução baseada em amostragem para reduzir dados multivariados em redes de sensores sem fio e os resultados das avaliações feitas com o algoritmo de amostragem proposto.

3 TRABALHOS RELACIONADOS

O presente trabalho propõe um algoritmo de amostragem para efetuar redução de dados multivariados em redes de sensores sem fio e utiliza técnicas de análise de componentes para auxiliar na redução desses dados. Dessa forma, é necessário discutir sobre os trabalhos existentes na literatura que empregam técnicas de redução de dados. Para efetuar essa redução diversas técnicas têm sido empregadas, dentre as quais é possível citar a agregação de dados, amostragem adaptativa, redução baseada em *stream* de dados e a redução de dados multivariados. Além disso, é preciso discutir sobre os trabalhos que utilizam análise de componentes em redes de sensores sem fio. Assim, na seção 3.1, serão apresentados os trabalhos referentes às técnicas de redução de dados univariados e, na seção 3.2, os trabalhos que empregam técnicas de redução de dados multivariados, incluindo alguns trabalhos que utilizam técnicas de análise de componentes em redes de sensores sem fio.

3.1 Redução de dados univariados

A maior parte dos trabalhos que utilizam técnicas de redução de dados em redes de sensores sem fio abordam a redução de dados univariados. Dessa forma, nesta seção, serão apresentadas as principais abordagens utilizadas para esse fim. Dentre as técnicas utilizadas para reduzir dados univariados em redes de sensores sem fio, destacam-se a técnica de agregação de dados, amostragem adaptativa e a redução baseada em *stream* de dados, apresentadas a seguir.

3.1.1 Agregação de dados

Uma das técnicas mais utilizadas para efetuar redução de dados univariados em redes de sensores sem fio é a agregação de dados. A utilização de agregação como técnica de redução visa combinar os dados vindos de diferentes fontes, eliminando redundância, minimizando o número de transmissões e economizando energia.

A agregação possibilita a obtenção de ganhos significativos, principalmente quando há um grande número de nós fonte e estes estão próximos entre si e a vários saltos do sorvedouro. Todavia, é preciso considerar o impacto da utilização dessa técnica em redes de sensores sem

fio (KRISHNAMACHARI; ESTRIN; WICKER, 2002). Uma consideração importante é que a utilização de esquemas de agregação ótimos é uma tarefa NP-Difícil, uma vez que a construção dessa árvore corresponde ao problema da Árvore de Steiner, comprovadamente um problema NP-Difícil (GAREY; JOHNSON, 1979). Isso torna necessária a aplicação de esquemas subótimos para realizar esse processo, tais como *Center at Nearest Source* - CNS, *Shortest Paths Tree* - SPT, *Greedy Incremental Tree* - GIT e *Information-Fusion-based Role Assignment* - InFRA (NAKAMURA et al., 2009). Além disso, ela pode proporcionar um aumento do atraso, que no pior caso será proporcional à distância entre o sorvedouro e a fonte mais distante. Outro fator importante a ser considerado é que seu desempenho é influenciado também pelo número de fontes, pela topologia e pela densidade da rede, o que poderia inviabilizar sua utilização em uma grande gama de aplicações em redes de sensores sem fio.

Diversas abordagens têm sido apresentadas na busca de melhorar os resultados obtidos com a agregação de dados. Uma dessas abordagens é seleção inteligente de árvores de agregação, a partir de um conjunto de árvores de agregação candidatas. Em (DASGUPTA; KALPAKIS; NAMJOSHI, 2003), é feita uma adaptação do algoritmo *Maximum Lifetime Data Aggregation* - MLDA (KALPAKIS; DASGUPTA; NAMJOSHI, 2002), uma vez que sua complexidade o torna inviável para redes de sensores de grande porte. O algoritmo utiliza uma seleção inteligente de árvores de agregação, a partir de um conjunto de árvores de agregação candidatas. A abordagem considera uma rede hierárquica, onde os dados coletados pelos nós em cada agrupamento são agregados e transmitidos ao sorvedouro através de um nó líder. Nas primeiras δ transmissões, o algoritmo utiliza δ árvores de agregação, e a partir daí, nas próximas transmissões, apenas algumas dessas árvores são selecionadas, possibilitando uma economia de energia.

Outra possibilidade referente ao tratamento das árvores de agregação é a utilização de algoritmos baseados em árvore de menor caminho (ZHANG et al., 2007). A ideia é ter um algoritmo com duas fases, sendo que na primeira ele constrói uma árvore de caminho de energia máxima, que balanceia o consumo de energia entre os nós. Na segunda fase, a árvore é reestruturada através de um algoritmo de menor caminho, para tentar minimizar a latência na agregação dos dados. O objetivo é encontrar uma árvore de agregação abrangendo todos os sensores na área monitorada e encontrar uma raiz adequada para a coleta de dados. Dessa forma, a árvore pode minimizar a latência na agregação e balancear o consumo de energia, maximizando o tempo de vida da rede. O interesse então é encontrar o menor caminho, com a maior quantidade de energia residual. Nesse caso, a agregação executa uma fusão na rede dos pacotes de dados vindos de diferentes sensores para a raiz, com o intuito de minimizar o número de transmissões e o tamanho dos dados, e, consequentemente, economizar energia.

Além das formas mencionadas, é possível ainda realizar a agregação de acordo com os

recursos disponíveis nos nós sensores. Em (ZHU; PAPAVASSILIOU, 2004), propõe-se um método que realiza agregação de dados local, considerando o compromisso entre latência, eficiência de energia e qualidade de acordo com os recursos e requisitos específicos da aplicação, reduzindo a quantidade de informações transmitidas na rede. Assim, quando uma estação base solicita uma tarefa, ela informa a todos os nós as restrições a serem consideradas. Cada nó da rede decide se participa ou não da tarefa, e aqueles que participam, de acordo com seu nível de energia e o tempo necessário para entrega dos dados, agregam os dados ou simplesmente retransmitem sem realizar a agregação.

Apesar de todas as melhorias efetuadas no processo de agregação dos dados, um problema ainda precisa ser tratado. Parte dos dados que seriam usados na agregação podem se perder, uma vez que a transmissão de dados em redes de sensores sem fio está sujeita a falhas e perda de pacotes, o que pode, consequentemente, comprometer a precisão dos dados. Essas perdas afetam ainda mais a precisão das informações quando ocorrem próximas à raiz. Muitos trabalhos tentam resolver esse problema através da retransmissão dos pacotes perdidos, o que pode, no entanto, elevar o tempo gasto na comunicação e o consumo de energia. Para reduzir o impacto desses problemas, em (ANISI; REZAZADEH; DEHGHAN, 2008), é proposta a abordagem chamada *Fault-tolerant Energy-efficient Data Aggregation* - FEDA. Nessa técnica, quando ocorrem perdas de pacotes entre dois nós, outro nó da rede que tenha recebido esse pacote agrega-o com seus próprios dados e os envia. A abordagem utiliza caminhos disponíveis redundantes, para minimizar a perda de pacotes e entregar o dado agregado corretamente ao sorvedouro.

Uma consideração importante a respeito desses trabalhos é que nenhum deles faz uma análise aprofundada sobre a precisão dos dados após a realização da agregação. Essa característica pode ser de fundamental importância em várias aplicações, merecendo então, ser melhor avaliada. Além disso, uma vez que são agregados os dados de diferentes nós, esses dados possuem características multivariadas e, dessa forma, pode ser importante a realização de uma análise multivariada nesses dados, que é o foco desta dissertação.

3.1.2 Amostragem adaptativa

Além da agregação de dados, outras abordagens têm sido apresentadas para efetuar a redução de dados univariados em redes de sensores sem fio. Uma dessas abordagens é a utilização da técnica de amostras adaptativas. Essa técnica modifica a forma de sensoriamento, de acordo com os requisitos da aplicação e os recursos dos nós, com o objetivo de propagar apenas uma amostra contendo a informação mais relevante para a aplicação.

As redes de sensores precisam ser autoconfiguráveis e a grande quantidade de dispositivos e a característica dinâmica dos ambientes são desafios para se obter essa autonomia. Quanto maior a variação no ambiente monitorado, menor tende a ser a taxa de precisão das informações. Devido a esses fatores, um mecanismo de controle é utilizado em cada sensor, o que faz com que a taxa de sensoriamento seja dinâmica e adaptativa (MARBINI; SACKS, 2003). Esse mecanismo trabalha com um modelo do ambiente a ser monitorado e, de acordo com esse modelo, um número maior ou menor de coletas será feito. A quantidade de amostras e a complexidade do modelo influenciam no uso de recursos como memória e energia, criando a necessidade de se obter um compromisso entre a taxa de amostragem e a precisão das informações.

Uma das formas de amostragem é a utilização de esquemas de predição dos valores coletados, nos quais apenas as informações que desviam dessa predição são transmitidas. Geralmente, esse processo pode ser divido em dois passos. No primeiro passo, uma estimativa inicial do ambiente é formada usando um subconjunto dos nós sensores. Esse subconjunto envia as informações para um centro de fusão que, com base nessa estimativa, envia mensagem para que outros sensores sejam ativados, afim de melhorar a precisão das informações, o que ocorre no segundo passo. A ideia principal é que com a estimativa inicial seja possível conhecer as correlações no ambiente, podendo assim, determinar o número de sensores que não precisam ser ativados (WILLETT; MARTIN; NOWAK, 2004). Fator relevante nesse cenário e não abordado no trabalho é que uma variação maior nos dados coletados pelos sensores pode afetar drasticamente o desempenho da técnica, uma vez que o número de sensores que serão ativados pelo centro de fusão pode ser muito grande.

Outra limitação a ser considerada é que a técnica de predição, apesar de reduzir o consumo de energia na rede, precisa ter um conhecimento "a priori" das informações, fazendo com que o modelo tenha que ser constantemente atualizado, aumentando, assim, os custos da comunicação. Uma maneira de tratar esse problema é o uso de esquemas de predição adaptativa não baseados em um modelo de conhecimento prévio (SANTINI; ROMER, 2006). Nesse caso, são usados filtros de predição no nó fonte e no sorvedouro e quando o nó coleta os dados e envia para o sorvedouro, ambos aplicam o algoritmo de predição e fazem uma estimativa de erro, que é comparada com um limite previamente definido. Assim, o nó enviará informações somente se esse limite de erro for excedido.

Em (SANTINI; ROMER, 2006), contudo, considerou-se apenas ambientes que necessitam de entrega contínua de dados em intervalos de tempo regulares, ou seja, as irregularidades temporais não foram consideradas. No entanto, considerar as irregularidades espaço-temporais no processo de amostragem em redes de sensores sem fio é muito importante, uma vez que os

fenômenos não são distribuídos uniformemente no espaço, pois os recursos dos sensores variam para executar o sensoriamento, e amostragem temporal regular requer relógios sincronizados em todos os pontos de mensuração, aumentando o custo de transmissão e o gasto de energia, o que pode inviabilizar sua aplicação (GANESAN et al., 2004). É importante destacar que, assim como no caso da agregação de dados, a utilização de amostragem adaptativa também pode considerar as características multivariadas dos dados, fator não abordado pelos trabalhos citados, diferentemente da abordagem empregada nesta dissertação.

3.1.3 Redução baseada em stream de dados

Outro método utilizado para efetuar a redução de dados em redes de sensores sem fio é a redução baseada em *stream* de dados. Em redes de sensores sem fio, alguns fenômenos monitorados geram dados ao longo do tempo com características que possibilitam classificá-los como *stream* de dados (AQUINO, 2008). O *stream*, nesse caso, representa as amostras da população de interesse, possui tamanho moderado, é impreciso e com ruído. Para o tratamento desse tipo de dados, podem ser utilizados algoritmos baseados em técnicas de *stream* de dados como amostragem, histograma, rascunho e janela deslizante.

Em (AQUINO et al., 2007), são apresentados dois algoritmos baseados em *stream* de dados para efetuar a redução de dados em redes de sensores sem fio. O primeiro algoritmo realiza um processo de amostragem baseada em um histograma do *stream* sensoriado. O segundo algoritmo utiliza rascunho, sendo também baseado em informações do histograma do *stream* de dados. Os dois algoritmos foram avaliados considerando uma rede plana, onde cada sensor coleta suas informações e efetua a redução dos dados para enviar ao sorvedouro. A avaliação foi feita em relação ao comportamento da rede e à qualidade dos dados reduzidos. No que se refere ao comportamento da rede, os dois algoritmos foram avaliados e, para avaliar a qualidade dos dados, apenas o algoritmo de amostragem foi utilizado. Resultados mostraram que o uso do algoritmo de amostragem reduz a quantidade de dados transmitidos, com reduzidas perdas de qualidade. Além disso, a utilização dos dois algoritmos resultaram em redução do consumo de energia e do atraso na rede.

Em (AQUINO et al., 2008), apresenta-se um algoritmo para redução baseada em *stream* em redes de sensores sem fio hierárquicas, onde a redução dos dados gerados pelos nós integrantes do agrupamento é efetuada pelo nó líder. O objetivo desse trabalho foi investigar se os algoritmos de redução baseada em *stream* de dados são eficientes para uma rede hierárquica. Para isso, as técnicas de amostragem e rascunho, utilizadas para redução considerando redes planas em (AQUINO et al., 2007), foram empregadas para a redução em redes hierárquicas. Resultados

mostram que as técnicas são eficientes também em redes hierárquicas, conseguindo reduzir o atraso e o consumo de energia, com reduzidas perdas na qualidade dos dados. Entretanto, como nos trabalhos mencionados anteriormente, a redução é efetuada considerando apenas dados univariados. Uma vez que diferentes fenômenos podem ser monitorados simultaneamente, ou ainda dados de diferentes nós podem ser reduzidos, por exemplo em uma rede hierárquica, através de um nó líder de agrupamento, os dados podem ser caracterizados como multivariados. Nesse caso, a utilização de técnicas de redução que consideram essas características, como a apresentada nesta dissertação, é importante para a obtenção de melhores resultados. Vale ressaltar que esses trabalhos foram utilizados como base/motivação para os estudos realizados nesta dissertação.

3.2 Redução de dados multivariados

Na redução de dados multivariados, alguns trabalhos utilizam métodos para estimar o comportamento do dado multivariado a ser sensoriado e apenas as diferenças observadas ao longo do tempo são enviadas. Em cada rede, dependendo da aplicação, os requisitos e as características dos dados podem ser diferentes, e as técnicas de redução de dados multivariados podem considerar esses aspectos. Diferentemente dessa abordagem, o presente trabalho não visa estimar o comportamento dos dados, mas efetuar uma amostra dos dados sensoriados com o auxílio de técnicas de análise multivariada.

Uma consideração importante é que grande parte dos trabalhos que utilizam técnicas de redução de dados em redes de sensores sem fio não abordam esse tipo de redução de dados. Entretanto, diversos métodos podem ser utilizados para esse fim. Em (SEO; KANG; RYU, 2005), é feita uma comparação entre os métodos *Discrete Wavelet Transformation* - DWT, *Hierarchical Clustering* - HCL, *Amostragem* e *Singular Value Decomposition* - SVD. Os métodos são avaliados variando o tamanho dos dados e o tipo de dado, utilizando conjuntos de dados reais e outros sintéticos. Segundo os autores, o método de *Amostragem* teve desempenho superior com a variação do tamanho e do tipo de dados. Nesse trabalho, diferentemente da técnica proposta nesta dissertação, para se determinar o erro entre o conjunto de dados original e o conjunto reduzido é feita uma reconstrução da matriz de dados original, e o erro é calculado com base nessa matriz reconstruída. Além disso, apesar do principal motivo de se aplicar a redução de dados em redes de sensores sem fio ser o problema de energia, não é feita nenhuma avaliação do consumo de energia aplicando as técnicas citadas. Essa avaliação é um dos focos nessa dissertação, que também realiza um processo de amostragem.

Outro método, também empregado na redução de dados multivariados em redes de sen-

sores sem fio é a ICA, apresentado na seção 2.2.2. Em (CVEJIC; BULL; CANAGARAJAH, 2007), apresenta-se um algoritmo para melhorar a fusão de imagens de vigilância, baseado em Análise de Componentes Independentes, onde a codificação esparsa dos coeficientes de ICA diminuem o ruído nas imagens fundidas. O objetivo da fusão de imagens, além de diminuir o tráfego de informações, é criar imagens mais perceptíveis e adequadas para um processamento posterior. No método utilizado, para realizar a redução, um pré-processamento é realizado com a técnica PCA, através da qual a dimensão dos dados é reduzida. Posteriormente, os vetores base são estatisticamente selecionados, através de ICA. O próximo passo é a aplicação de um algoritmo para reconstrução das imagens, em conjunto com um esquema para reduzir o ruído. Resultados apresentados mostram que o método proposto é eficiente em termos da qualidade das imagens após a fusão e também da redução de ruído. O cenário abordado nesse trabalho, bem como as métricas utilizadas para avaliação do algoritmo são diferentes do que é empregado nessa dissertação. As características dos dados das aplicações para tratamento de imagens são diferentes das características dos dados gerados nas aplicações gerais de monitoramento e, devido a isso, as métricas de avaliação também diferem, como é o caso da avaliação do ruído nas imagens, o que não é avaliado nas aplicações gerais de monitoramento.

Além das técnicas mencionadas, outra técnica que pode ser empregada é a Análise de Componentes Principais (PCA), apresentada na seção 2.2.1. Ainda poucos trabalhos têm sido apresentados usando PCA para reduzir dados em redes de sensores sem fio. Nos trabalhos existentes, PCA tem sido avaliado, geralmente, em aplicações específicas. Um exemplo de aplicação é o monitoramento de dados sensoriados sobre vibração em larga escala para sistemas de monitoramento estrutural (LI; ZHANG, 2006). Nesse caso, utilizou-se um algoritmo que integra um método de compressão de dados baseado em PCA com o monitoramento das características do ambiente para realizar a predição e reduzir a quantidade de informações transmitidas. Resultados mostram que a técnica consegue diminuir a quantidade de dados com reduzidas perdas de informação, não citando, entretanto, valores para essas perdas. Além disso, o emprego de PCA se dá em conjunto com uma técnica de predição. Embora essa técnica de predição não seja apresentada em detalhes no trabalho, é possível que a mesma gere um custo de comunicação maior na rede, conforme citado por (SANTINI; ROMER, 2006).

Outro cenário em que PCA tem sido utilizado em condições específicas, considera seu emprego em aplicações nas quais as características de tempo e espaço são estacionárias, como por exemplo, a codificação de audio (ROY; VETTERLI, 2008). Nesse trabalho, propõe-se uma abordagem baseada em transformação para reduzir a quantidade de dados enviados para um centro de fusão. Nessa técnica, cada sensor aplica uma transformação linear em seus dados através de PCA, para que seja enviada para o centro de fusão uma quantidade reduzida, a fim

de minimizar o consumo de energia na rede. No centro de fusão, o sinal é reconstruído e uma análise da distorção desse sinal é efetuada através do erro médio quadrático *Mean Squared Error* - MSE. Os dados seguem uma distribuição *normal* e utiliza o modelo de correlação Gauss-Markov de primeira ordem. Resultados mostram que a técnica obteve bons resultados, garantindo um nível de precisão adequado. Contudo, a avaliação é apenas teórica e com um foco diferente do abordado nessa dissertação, considerando aplicações e análises diferentes.

Além do emprego para efetuar redução de dados, diminuindo o consumo de energia na rede, PCA pode ser utilizado também para auxiliar na deteção de anomalias nos dados coletados por diferentes sensores (CHATZIGIANNAKIS; PAPAVASSILIOU, 2007). Nesse caso, PCA é utilizado com o propósito de reduzir a quantidade de dados, mantendo sua qualidade, de forma a facilitar a aplicação de testes nos dados para encontrar essas anomalias. Para efetuar essa redução, algumas componentes principais são utilizadas para selecionar uma amostra com os dados que cobrem maior parte da variância do conjunto de dados originais. Após a seleção desses dados feita através de PCA, o teste *Squared Prediction Error* - SPE é empregado para encontrar as anomalias. Quando o limite previamente estabelecido para esse erro é excedido, indica que está ocorrendo uma anomalia, e é possível determinar os sensores responsáveis por essas informações anômalas. A avaliação do método é feita utilizando-se dados meteorológicos reais e inserindo-se anomalias nesses dados aleatoriamente, para então medir sua capacidade de deteção. Fator interessante no trabalho é a utilização da matriz de correlação ao invés da matriz de covariância, mostrando que ambas podem ser empregadas e que podem obter resultados diferentes. Outro fator que merece destaque é a seleção dos dados utilizando-se um número de componentes principais variável, o que pode ser objeto de um estudo futuro utilizando o algoritmo proposto nessa dissertação.

É importante ressaltar que grande parte dos trabalhos que empregam PCA, efetuam redução de dimensionalidade dos dados, necessitando, posteriormente, de reconstruir os dados originais para então analisar os resultados. Diferentemente desses trabalhos, o algoritmo proposto nessa dissertação não efetua redução de dimensionalidade dos dados, mas sim utiliza técnicas de análise de componentes para classificar os dados e auxiliar no processo de amostragem. Com isso, não é necessária a reconstrução dos dados originais para que possam ser feitas análises nesses dados e essas análises são feitas utilizando técnicas diferentes das utilizadas nesses trabalhos.

3.3 Conclusões parciais

Apesar das diversas soluções propostas até o momento, a redução de dados em redes de sensores sem fio ainda é um grande desafio. Problemas como economia de energia e atraso precisam de soluções mais eficientes para uma enorme gama de aplicações. Principalmente no que se refere à redução de dados multivariados ainda há uma significativa carência de soluções e a qualidade dos dados reduzidos é fator primordial de consideração. Além disso, chama a atenção o fato de que grande parte dos trabalhos que apresentam abordagens para redução de dados não fazem uma análise mais aprofundada da qualidade dos dados após efetuar a redução. Isso é muito importante, pelo fato de que em diversas aplicações, perdas na qualidade dos dados podem inviabilizar o emprego de determinadas soluções, embora essas soluções consigam reduzir o consumo de energia e o atraso, por exemplo. Nesse caso, a extensão do tempo de vida da rede ficaria em segundo plano.

Dentre as possíveis abordagens para redução de dados multivariados, a utilização de técnicas baseadas em análise de componentes se mostra viável. Essa viabilidade se deve aos resultados obtidos com o emprego de tais técnicas em redes de sensores e em outras áreas de aplicação, principalmente devido à possibilidade de efetuar a redução, mantendo a qualidade dos dados, como será demonstrado neste trabalho.

4 ALGORITMO DE AMOSTRAGEM MULTIVARIADA

Este capítulo apresenta uma solução baseada em análise de componentes para a redução de dados multivariados em redes de sensores sem fio, que utiliza amostragem. No entanto, antes de apresentar o algoritmo de redução a ser utilizado, o MuSA - **Mu***ltivariate* **S***ampling* **A***lgorithm*, algumas considerações se fazem necessárias para o emprego dessa técnica de redução de dados, e serão discutidas a seguir.

4.1 Contextualização

Para se projetar soluções eficientes para o problema de redução de dados em redes de sensores sem fio, são necessárias algumas considerações. Inicialmente, é preciso determinar as situações nas quais é possível efetuar a redução. Em redes de sensores sem fio, a redução de dados Ψ pode ser aplicada em diferentes situações, como por exemplo, no momento do sensoriamento e no nó líder de agrupamento. Essas situações são observadas tanto na redução de dados multivariados quanto univariados.

O primeiro caso refere-se à redução Ψ local, ou seja, no momento em que o dado é sensoriado, como ilustrado na figura 5. Nesse caso, cada nó possui sensores que monitoram diferentes fenômenos. No que se refere à redução Ψ no sensoriamento, existem diversos nós sensores s que coletam simultaneamente diferentes dados do ambiente monitorado, de tal forma que compõe-se $\mathcal{V}_n^s$, sendo s a quantidade de sensores por nó e n as diferentes leituras ao longo do tempo. Tais nós sensores podem efetuar a redução dos dados Ψ após diferentes n leituras do ambiente, evitando que dados desnecessários trafeguem na rede. A redução no momento do sensoriamento pode ocorrer caso a aplicação necessite de muitas leituras do ambiente em um curto período de tempo, o que torna necessário o processamento dos dados $\mathcal{V}_n^s$ através de Ψ, gerando uma amostra $\mathcal{V}_{n'}^{\prime s}$, onde $n' < n$. Para essa redução, é preciso garantir que as decisões D' tomadas pela aplicação não sejam comprometidas. Para isso serão utilizadas as regras $\mathcal{R}'_\Phi$ e $\mathcal{R}'_\Upsilon$.

O segundo caso refere-se à redução localizada ou redução no nó líder, mostrada na figura 6. Essa opção considera a utilização de um nó líder para concentrar os dados $\mathcal{V}$ lidos por um conjunto restrito de nós que monitoram o mesmo fenômeno, agrupados de acordo com algum

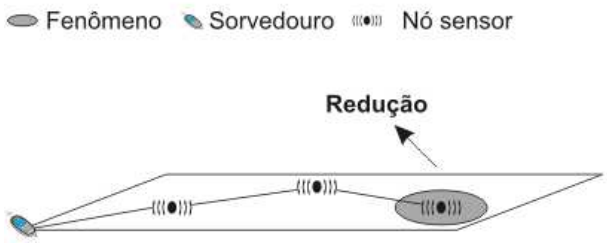


FIGURA 5: Redução no sensoriamento

interesse da aplicação. Nessa direção, a redução Ψ é efetuada no nó líder, permitindo que apenas os dados $\mathcal{V}'$ mais relevantes saiam do agrupamento de nós. Cada agrupamento possui um conjunto de sensores que reportam para o nó líder algumas leituras do ambiente. Com isso, esse nó será responsável por processar os dados sensoriados pelos nós em seu agrupamento. Note que cada nó i gera $\mathcal{V}_n^i$ dados e, ao chegar ao nó líder, tem-se $\mathcal{V} = (\mathcal{V}_n^1, \mathcal{V}_n^2, \ldots \mathcal{V}_n^s)$, onde s é o conjunto de sensores do agrupamento. Nesse caso, o conjunto de dados $\mathcal{V}_n^s$ também é processado por Ψ, gerando um conjunto reduzido $\mathcal{V}_{n'}^{\prime s}$, com $n' < n$. Como no caso anterior, é necessário garantir que as decisões D' tomadas pela aplicação não sejam comprometidas, através das regras $\mathcal{R}'_\Phi$ e $\mathcal{R}'_\Upsilon$.

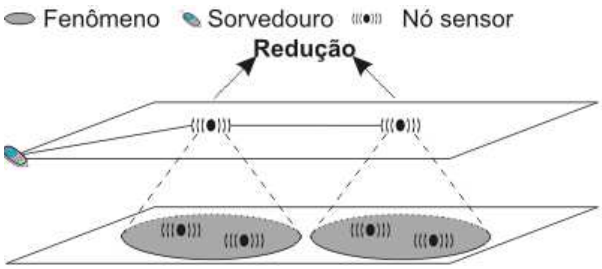


FIGURA 6: Redução no nó líder

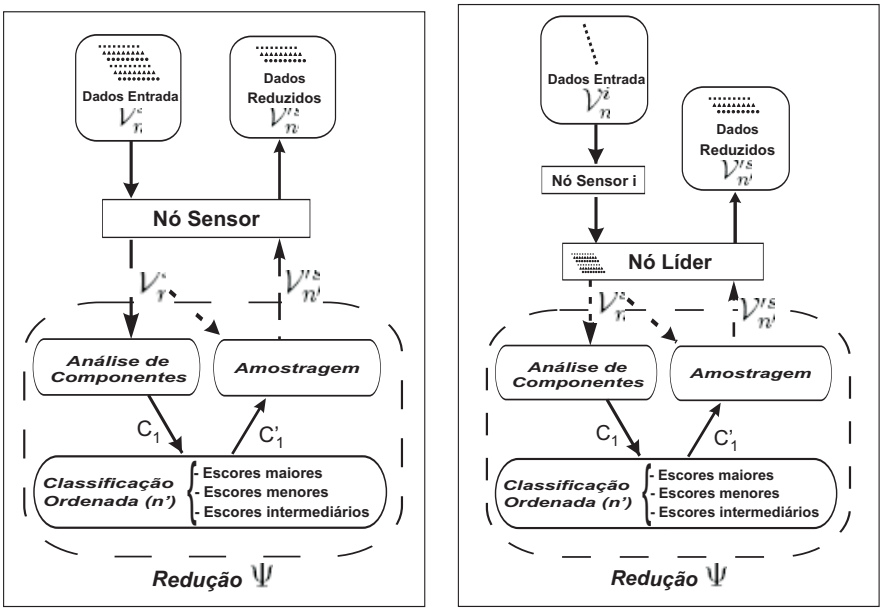
Outra consideração importante para a utilização de técnicas de redução diz respeito à representatividade dos dados reduzidos, ou seja, uma comparação entre os dados reduzidos e os dados originais. É evidente que cada aplicação tem seus requisitos de qualidade, e assim, para cada aplicação, diferentes métricas de avaliação podem ser empregadas. Neste trabalho, as decisões de aceitação dos dados são tomadas a partir das regras $\mathcal{R}'_\Phi$ e $\mathcal{R}'_\Upsilon$, definidas na seção 2.3.

Por fim, é preciso analisar qual o nível adequado de redução suportado nas aplicações, de tal forma que essa redução não comprometa as decisões D' a serem tomadas. De acordo com as decisões exigidas pela aplicação, pode-se estipular um nível adequado para essa redução Ψ, ou seja, o quão pequeno será n' em $\mathcal{V}_{n'}^{\prime s}$. Entretanto, não é objetivo deste trabalho determinar qual é esse nível adequado para cada tipo de aplicação, especialmente devido ao fato de que diversas aplicações teriam de ser abordados, bem como inúmeros níveis de redução, o

que poderia tornar o estudo inviável. Além disso, para um mesmo tipo de aplicação, valores diferentes podem ser considerados adequados, em função do domínio em que estão inseridas. Neste trabalho, foi estipulado um nível máximo de redução $(n' = \log_2 n)$ e observado como os algoritmos se comportam. Partindo desse nível, o projetista da aplicação poderá determinar o quão pequeno pode ser n'.

4.2 Amostragem baseada em análise de componentes

Com base nas considerações apresentadas na seção anterior, descreve-se então uma proposta de solução para a redução de dados multivariados em redes de sensores sem fio. A solução, mostrada na figura 7, é direcionada para aplicações gerais de monitoramento em redes de sensores sem fio que geram dados multivariados e utiliza um algoritmo de amostragem para efetuar a redução Ψ.



(a) Redução no sensoriamento (b) Redução no nó líder

FIGURA 7: Amostragem baseada em análise de componentes para redução de dados multivariados em redes de sensores sem fio

Para que a redução Ψ seja efetuada no sensoriamento (figura 7(a)), o nó sensor coleta as informações do ambiente e ele mesmo irá efetuar a redução Ψ. Nesse caso, diferentes fenômenos são monitorados simultaneamente por cada nó da rede, formando localmente o conjunto multivariado $\mathcal{V}_n^s$. Para efetuar Ψ, primeiramente é feito um processamento nos dados sensoriados. Esse processamento é feito utilizando uma técnica de análise de componentes (neste trabalho utilizou-se as técnicas PCA, PCA-robusta e ICA), com o objetivo de criar uma

classificação ordenada dos dados coletados por cada sensor e, com isso, fazer com que as decisões sobre os dados possam ser tomadas adequadamente ao se montar o conjunto de dados reduzido $\mathcal{V}'$. Essa classificação é feita com base no número de dados n' do conjunto reduzido e no tipo de escore a ser utilizado, ou seja, os escores maiores, menores ou intermediários da primeira componente. A utilização da técnica de análise de componentes para criar a classificação dos dados, bem como o processo de amostragem serão discutidos em detalhes na próxima seção, onde será apresentado o algoritmo proposto neste trabalho. Após o processamento dos dados, realiza-se então uma classificação ordenada dos dados sensoriados, para identificar os dados mais relevantes para a aplicação e, posteriormente, é executada a amostragem, que, de acordo com o nível de redução desejado, irá selecionar uma amostra dos dados coletados, contendo os dados identificados como mais relevantes para a aplicação, fazendo com que as decisões tomadas pela aplicação não sejam comprometidas. Terminado esse processo de amostragem, conclui-se a tarefa de redução dos dados e o conjunto de dados reduzidos $\mathcal{V}'$ é então enviado pelo nó sensor até o sorvedouro.

Para aplicar Ψ no nó líder, o processo é semelhante ao executado na redução no sensoriamento, conforme ilustrado na figura 7(b). A única diferença é que, nesse cenário, o nó sensor i coleta as informações $\mathcal{V}_n^i$ e as envia para o nó líder, que nesse caso é o responsável por receber os dados coletados pelos diversos nós i do agrupamento e efetuar a redução em $\mathcal{V}_n^s$. Nesse caso, todos os nós do agrupamento monitoram um único fenômeno e transmitem os dados coletados ao nó líder. Após receber os dados coletados, o nó líder realiza o processo de redução Ψ de forma idêntica ao apresentado anteriormente. Ao final do processo de redução dos dados, o nó líder envia o conjunto de dados reduzido $\mathcal{V}'$ para o sorvedouro.

4.3 Algoritmo MuSA

Esta seção apresenta o algoritmo de amostragem baseado em análise de componentes para redução de dados multivariados em redes de sensores sem fio, o MuSA - **Mu***ltivariate* **S***ampling* **A***lgorithm*. O principal objetivo desse algoritmo é reduzir redundâncias e detalhes pouco significativos, gerando uma nova coleção de dados $\mathcal{V}'$, nesse caso uma amostra dos dados originais $\mathcal{V}$, que represente as características do conjunto de dados original com a perda mínima de informação. Ou seja, após efetuar Ψ em $\mathcal{V}$ através do MuSA, o conjunto de decisões D' tomadas deve ser correspondente ao conjunto de decisões D que seriam tomadas caso todo o conjunto $\mathcal{V}$ fosse utilizado.

Diferentemente da maioria dos trabalhos que empregam técnicas de análise de componentes, o algoritmo proposto neste trabalho não realiza redução de dimensionalidade. O algoritmo

proposto utiliza essas técnicas para classificar os dados multivariados e, com isso, identificar o que pode ser reduzido sem comprometer a representatividade dos dados. Como mencionado no capítulo 3, a primeira componente principal C_1 contém a maior variância entre todas as componentes principais. Além disso, após o cálculo das componentes principais C, é possível determinar seus valores numéricos (ou escores). Esses escores indicam a participação de cada elemento da componente principal na variância global dessa componente. Determinados os escores, é possível realizar análises estatísticas nos mesmos, como análise de variância, análise de regressão, entre outras. Com isso, é possível saber os escores mais representativos de uma componente. Essas características são consideradas também para as componentes independentes.

Conforme apresentado na formulação do problema, na seção 1.3, considere a matriz com dados de entrada $\mathcal{V}_n^s$, o número de componentes C dessa matriz é igual a s e o número de escores em cada componente é n. Dessa forma, cada escore de uma componente está relacionado a um dado da coluna correspondente da matriz de entrada $\mathcal{V}$. Ou seja, cada escore de uma componente está relacionado a uma linha da respectiva coluna da matriz de entrada. Como é possível determinar os escores mais representativos para uma componente e esse escore está relacionado com determinado dado (linha de determinada coluna da matriz de entrada), é possível, então, determinar as linhas da matriz original $\mathcal{V}$ que são mais relevantes, que nesse caso são aquelas que cobrem maior variância dos dados.

Uma vez que a primeira componente C_1 explica a maior variância dos dados, a classificação desses dados, pelo algoritmo MuSA é feita a partir dessa componente. MuSA realiza o processo de amostragem referenciando cada escore de C_1 com uma linha de $\mathcal{V}$. Sendo assim, a seleção dos dados que irão compor $\mathcal{V}'$ pode ser efetuada de três formas: considerando os escores maiores, menores ou intermediários dessa componente. Através da realização de análise de variância nos escores de C_1, verificou-se que em ambos os casos, é possível obter uma variância satisfatória. Vale ressaltar que, para diferentes aplicações, modificações no funcionamento do algoritmo podem ser facilmente implementadas, por exemplo na forma de seleção da amostra, possibilitando uma melhora no desempenho do mesmo. Outro ponto importante da técnica proposta é que não há a necessidade de reconstrução dos dados após sua chegada ao sorvedouro, uma vez que não é feita redução de dimensionalidade.

Além disso, como foi mencionado anteriormente, é preciso analisar qual o nível adequado de redução suportado nas aplicações, de tal forma que essa redução não comprometa as decisões D' tomadas sobre os dados. Uma vez que as aplicações possuem características e necessidades diferentes, para cada aplicação de interesse, diferentes níveis de redução devem ser aplicados e analisados, a fim de se determinar o nível mais adequado. Com isso, utilizou-

se dois níveis arbitrários de redução $n' = n/2$ e $n' = \log_2 n$, com o intuito de introduzir possíveis limites de redução para as aplicações. Dessa forma, dependendo da aplicação a ser considerada, diferentes valores devem ser aplicados e analisados.

4.3.1 Funcionamento básico do MuSA

Para ilustrar os dados multivariados gerados nas aplicações, considere novamente a matriz $\mathcal{V}_n^s$ os dados de entrada, onde $n > 0$ representa os valores monitorados por cada sensor e $s \geq 1$ representa os sensores responsáveis por obter informações do ambiente. Com isso, para descrever o funcionamento básico do algoritmo MuSA para aplicar Ψ em $\mathcal{V}$, é possível considerar os seguintes passos, ilustrados na figura 8.

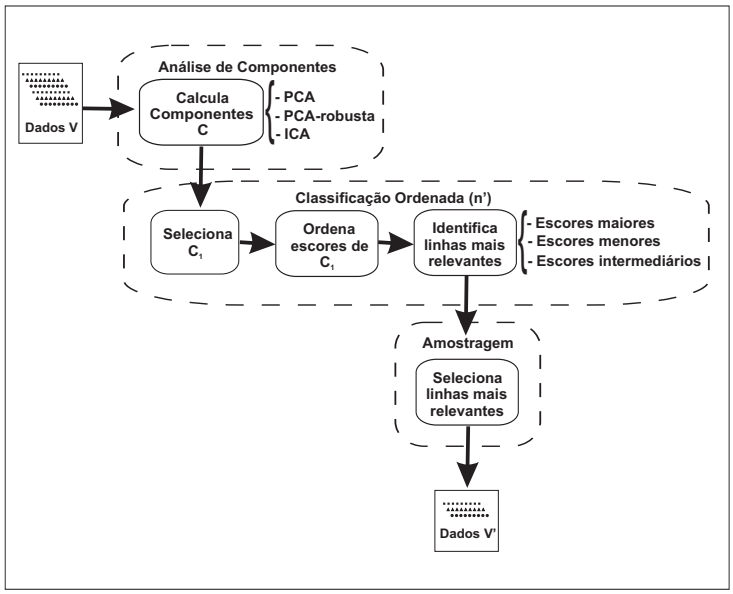


FIGURA 8: Passos do algoritmo MuSA

Primeiramente, a técnica para análise de componentes escolhida é utilizada para calcular as componentes C do conjunto original de dados sensoriados $\mathcal{V}$. Após o cálculo das componentes, a primeira componente C_1 é selecionada e seus respectivos escores são ordenados. Com isso, em função do tipo de escores a ser utilizado, ou seja, os escores maiores, menores ou intermediários, as posições desses escores da componente C_1 são usadas para referenciar as posições das linhas em $\mathcal{V}$ que irão compor o conjunto de dados reduzido $\mathcal{V}'$, de acordo com o nível de redução n' empregado. Por fim, o conjunto de dados reduzido $\mathcal{V}'$, contendo as linhas de $\mathcal{V}$ mais representativas para a aplicação, é obtido e posteriormente enviado ao sorvedouro. O pseudo-código é mostrado no algoritmo 1.

- Na linha 2, tem-se o cálculo das componentes, através da técnica escolhida. A or-

Algoritmo 1: Redução multivariada através do MuSA

Entrada: $\mathcal{V}$ – dados de entrada
Entrada: tec – técnica de análise de componentes
Entrada: n' – tamanho da redução
Entrada: esc – escore a ser utilizado
Saída: $\mathcal{V}'$ – dados reduzidos

```
1 início
2 |  C ← calculaComponentes(V, tec)
3 |  C_1 ← a primeira componente de C
4 |  I ← ordena(C_1)  /* Ordena escores de C_1 */
5 |  I ← ordena(I, n', esc)  /* I contém os escores de C_1 */
6 |  para i ← 1 até n' faça
7 |  |  V'_i ← V'_{I_i}
8 |  fim
9 fim
```

dem de complexidade do cálculo de PCA pode ser estimada em $O(s^2s' + s^2n)$, onde s corresponde ao número de sensores do conjunto de dados original, s' é o número de sensores do conjunto de dados reduzido e n o tamanho da amostra dos dados. Como nesse caso $s = s'$ e $s < n$, tem-se a ordem de complexidade $O(s^2n)$. Para o cálculo de ICA, considerando o algoritmo FastICA, essa ordem pode ser estimada em $O(sn)$ (ZARZOSO; COMON; KALLEL, 2006). A ordem de complexidade de PCA-robusta pode ser estimada em $O(sk^2n)$ (TORRE; BLACK, 2001), onde k representa o número de componentes principais desejado. Como neste trabalho, apenas a primeira componente principal é necessária, a ordem de complexidade é $O(sn)$.

- Na linha 3, a primeira componente C_1 é selecionada.
- Na linha 4, tem-se a ordenação do vetor com a primeira componente C_1, no qual os índices I dos escores de C_1 são obtidos. A ordem de complexidade da ordenação é $O(n \log_2 n)$, uma vez que $|C_1| = n$.
- Na linha 5, tem-se a ordenação do vetor I, em função do tipo de escore escolhido, considerando apenas os n' primeiros índices. Nesse caso, n' representa a quantidade de dados que irá compor o conjunto de dados reduzido $\mathcal{V}'$ e I contém os índices dos valores mais relevantes, ou seja, das linhas que contêm os dados mais significativos de $\mathcal{V}$. Isso é necessário para se manter a ordem de chegada dos elementos escolhidos para $\mathcal{V}'$. A ordem de complexidade da ordenação é $O(n' \log_2 n')$.
- Nas linhas 6 – 8, tem-se a montagem dos dados de saída reduzidos $\mathcal{V}'$, cuja ordem de complexidade é $O(n')$.

Sendo assim, no caso da utilização de PCA, a complexidade de tempo total é $O(s^2n) + O(n\log_2 n) + O(n'\log_2 n') + O(n') = O(s^2n) + O(n\log_2 n)$. Utilizando ICA ou PCA-robusta, a complexidade de tempo total é $O(sn) + O(n\log_2 n) + O(n'\log_2 n') + O(n') = O(sn) + O(n\log_2 n)$.

Para a complexidade de espaço, considere as matrizes $\mathcal{V}$, $\mathcal{V}'$, C, que correspondem respectivamente aos dados de entrada, dados de saída e componentes principais ou independentes. A complexidade de espaço é dada por $2\,O(s\,n) + O(s\,n') = O(s\,n)$. Uma vez que cada nó fonte envia $\mathcal{V}'$ até o sorvedouro, a complexidade de comunicação é $O(s\,n'\,\rho)$, onde ρ é a maior rota na rede.

A figura 9 ilustra um exemplo de execução do MuSA. Primeiramente tem-se uma matriz com os dados de entrada. Aplica-se então a técnica PCA, que determina as componentes principais. A primeira componente principal é selecionada e seus escores são ordenados. Cada um desses escores está relacionado a uma linha da matriz de entrada, ou seja, o escore 1 está relacionado à linha 1, o escore 2 à linha 2 e assim sucessivamente. Considerando que os escores a serem utilizados para realizar a classificação dos dados são os escores superiores, a ordem desses escores seria 3, 1, 4 e 2. A partir dessa ordenação, para reduzir os dados pela metade, são considerados como mais relevantes os escores 3 e 1. Dessa forma, são selecionadas as linhas 3 e 1 da matriz de entrada e essas linhas irão compor o conjunto reduzido. A execução utilizando PCA-robusta e ICA se dá da mesma forma.

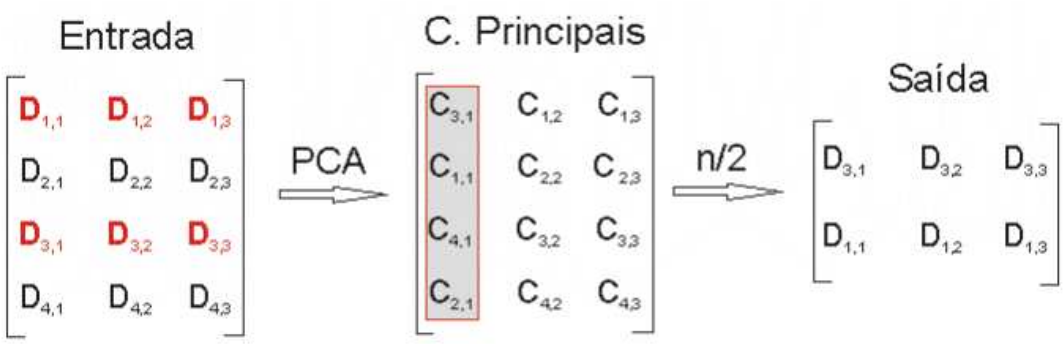


FIGURA 9: Exemplo de execução do algoritmo MuSA

4.4 Conclusões parciais

O presente capítulo apresentou solução proposta para redução de dados multivariados em redes de sensores sem fio, utilizando um algoritmo de amostragem baseado em análise

de componentes. O MuSA utiliza os escores da primeira componente C_1 obtida através da técnica de análise de componentes empregada para classificar os dados sensoriados e selecionar, a partir dessa classificação, uma amostra contendo os dados mais relevantes para a aplicação. De acordo com as características das aplicações nas quais será utilizado o algoritmo proposto, pode haver a necessidade de modificar seu funcionamento básico, bem como de analisar diferentes níveis de redução, para que o mesmo obtenha o melhor desempenho para a aplicação em questão. Nos próximos capítulos, serão apresentados os resultados de simulações com o algoritmo proposto.

5 REPRESENTATIVIDADE DOS DADOS

Este capítulo apresenta os resultados das simulações realizadas com o algoritmo descrito no capítulo anterior, com o intuito de avaliar a representatividade dos dados reduzidos em relação aos dados originais, representando as regras $\mathcal{R}'$, descritas na formulação do problema, na seção 1.3. Dessa forma, serão apresentados os cenários considerados nas simulações, a avaliação do algoritmo através de simulações utilizando PCA, ICA e PCA-robusta, considerando conjuntos de dados sintéticos e, posteriormente, apresenta-se a avaliação com conjuntos de dados reais e as conclusões obtidas a partir das simulações realizadas. Neste capítulo, serão apresentados apenas os resultados referentes à geração dos dados com a distribuição *normal*. Os resultados com a distribuição *skew-normal* serão discutidos no Apêndice, uma vez que os resultados observados são semelhantes aos da distribuição *normal*.

5.1 Cenários de simulação

Nesta seção, são apresentados os cenários utilizados para as simulações com o algoritmo MuSA empregando cada uma das técnicas de análise de componentes descritas anteriormente. Considerou-se para as avaliações a simulação de Monte Carlo (BUSTOS; FRERY, 1992), que é um método estatístico utilizado para a realização de simulações estocásticas. Assim, executou-se o algoritmo fazendo um cruzamento de todas as quantidades de sensores e de tamanho dos dados considerados nesse trabalho. Ou seja, para cada quantidade de sensores avaliada, simulou-se o algoritmo com todos os tamanhos de dados utilizados. Muito útil no processo de tomada de decisões, a simulação de Monte Carlo provê alta credibilidade, pelo fato de possibilitar a enumeração de uma grande variedade de resultados possíveis e suas probabilidades de ocorrência, evitando que decisões erradas, baseadas em resultados simplificados, possam ser tomadas.

Para a avaliação da qualidade dos dados, foram considerados conjuntos de dados sintéticos e outros pseudo-reais (ALBUQUERQUE, 2007), simulando as situações em que a redução é aplicada no sensoriamento e no nó líder. No caso da redução no nó líder, considera-se a redução em uma rede hierárquica, sendo todos os nós pertencentes ao mesmo agrupamento. Essa avaliação é feita através de simulações, com os algoritmos implementados através do programa estatístico R. O número de simulações necessárias para a avaliação foi calculado de

acordo com Jain (JAIN, 1991), e é dado por

$$rounds = \left(\frac{100\, z\, s_d}{p\, \overline{\mathcal{V}}}\right)^2, \tag{5.1}$$

onde z é uma constante de valor $1,96$, s_d é o desvio encontrado nas primeiras simulações, $\overline{\mathcal{V}}$ é a média dos valores obtidos e p é a porcentagem da média que se deseja obter como desvio. Nesse caso, a porcentagem utilizada foi de 5%, para que os resultados estivessem dentro de um intervalo satisfatório. Dessa forma, para cada cenário simulado realizou-se uma execução com 93 conjuntos de dados diferentes. Além disso, o algoritmo utiliza os escores intermediários, uma vez que os resultados da análise de variância são satisfatórios e por ser a situação em que o algoritmo obtém um menor erro absoluto relativo para os cenários avaliados.

5.2 Dados sintéticos

Para realizar a avaliação da representatividade dos dados utilizando dados sintéticos, empregou-se o MuSA no momento do sensoriamento e no nó líder, para comparar o desempenho do algoritmo com cada uma das três técnicas de análise de componentes citadas anteriormente. Para cada um desses cenários, a avaliação é feita de duas formas: (i) avalia-se o algoritmo a partir de dados gerados sem a presença de ruído; (ii) é feita uma análise com a inserção de ruído na geração dos dados de entrada, conforme descrito no capítulo 2. No caso da avaliação de $\mathcal{R}'_{\Upsilon}$, os resultados são apresentados com um intervalo de confiança de 95%.

5.2.1 Redução no sensoriamento

5.2.1.1 Geração dos dados sem ruído

O primeiro cenário a ser avaliado considera a redução no momento do sensoriamento, cuja geração dos dados é realizada sem a presença de ruído. Nesse cenário, cada nó da rede possui um arranjo de sensores que armazenam diferentes leituras e antes de propagá-las até o sorvedouro, é feita uma redução Ψ baseada na análise de componentes das diferentes variáveis sensoriadas. Para isso, fixou-se o número de sensores em um nó da rede em $s = 5$, simulando o sensoriamento de cinco variáveis diferentes. Dessa forma, a geração dos dados com distribuição *normal* foi feita a partir de cinco médias arbitrárias escolhidas para simular o sensoriamento das cinco diferentes variáveis. Os valores utilizados para as médias foram $\mu = \{10, 30, 50, 70, 90\}$, com um desvio padrão $\sigma = 10\%$ da média μ. O tamanho dos dados gerados foi variado em $n = \{256, 512,\ 1024, 2048\}$, definindo assim o $\mathcal{V}_n^s$ e aplicou-se as

reduções $n/2$ e $\log_2 n$. Nesse caso, os dados reduzidos serão $\mathcal{V}'^{s}_{n/2}$ e $\mathcal{V}'^{s}_{\log_2 n}$. É importante destacar que de acordo com a aplicação e os cenários considerados, outros valores para essa redução devem ser analisados, a fim de se determinar o mais adequado.

A primeira análise de representatividade dos dados considera $\mathcal{R}'_{\Phi}$, que representa os resultados obtidos pelo teste *ANOVA*, descrito na seção 2.3. Para essa análise, valores acima de $0,05$ são satisfatórios para a aceitação da hipótese nula, ou seja, indicam que não existem diferenças estatisticamente significativas entre os conjuntos de dados avaliados. Os valores observados para ambas as técnicas são apresentados na tabela 1. Esses resultados indicam que não existem diferenças significativas entre as variâncias dos conjuntos de dados original e reduzidos, tanto com a utilização da amostragem baseada em PCA quanto em ICA e PCA-robusta. Isso quer dizer que no conjunto de dados reduzidos $\mathcal{V}'$ estão representadas, de forma satisfatória, as variâncias existentes no conjunto de dados original. Logo, as decisões D' referentes a $\mathcal{R}'_{\Phi}$ podem ser tomadas.

TABELA 1: Análise da variância utilizando dados sintéticos sem ruído com redução no sensoriamento

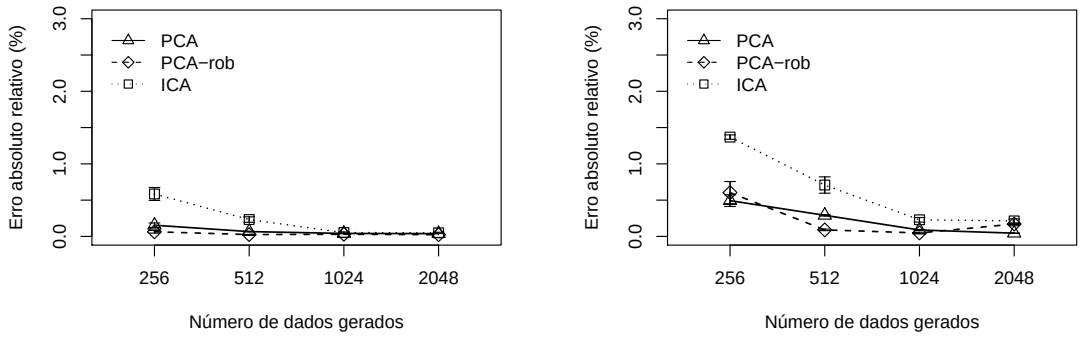
Distribuição	$(n = 256)$		$(n = 512)$		$(n = 1024)$		$(n = 2048)$	
normal	$n/2$	$\log_2 n$	$n/2$	$\log_2 n$	$n/2$	$\log_2 n$	$n/2$	$\log_2 n$
PCA	0,98	0,86	0,98	0,86	0,98	0,87	0,98	0,87
ICA	0,89	0,83	0,84	0,83	0,76	0,82	0,76	0,83
PCA-rob	0,98	0,85	0,98	0,86	0,98	0,86	0,98	0,87

A segunda análise considera o $\mathcal{R}'_{\Upsilon}$, que representa o erro absoluto relativo, descrito no capítulo 2. Resultados considerando o monitoramento realizado por 5 sensores podem ser observados na figura 10.

Considerando a redução $n/2$, mostrada na figura 10(a), os resultados obtidos com as três técnicas utilizadas foram muito satisfatórios, com $\mathcal{R}'_{\Upsilon}$ próximo de zero em todos os tamanhos de dados gerados. É importante destacar ainda que à medida em que aumenta-se a quantidade de dados gerados $\mathcal{V}$, em praticamente todos os casos, o $\mathcal{R}'_{\Upsilon}$ diminui em ambas as técnicas, o que ocorre devido ao fato de que uma maior quantidade de dados é gerada para cada fenômeno, considerando as mesmas médias μ e desvio padrão σ. Nesse cenário, os valores mais significativos foram observados com a técnica PCA-robusta e os menos significativos com a técnica ICA. Isso pode ser explicado pelo fato de que, conforme mencionado na seção 2.2.2, a primeira componente independente pode não possuir a maior porcentagem de variância dos dados, fazendo com que o processo de amostragem seja menos eficiente.

No que se refere à redução $\log_2 n$, mostrada na figura 10(b), novamente o MuSA foi bastante satisfatório com as três técnicas de análise de componentes utilizadas. Mais uma

vez, os valores observados para o $\mathcal{R}'_{\Upsilon}$ foram muito baixos em todos os casos, sendo o maior valor encontrado para a amostragem baseada em ICA, quando gerou-se $\mathcal{V}^{5}_{256}$ dados, ou seja, uma matriz com 5 colunas e 256 linhas. Nesse caso, o $\mathcal{R}'_{\Upsilon}$ foi igual a $1,3\%$. Além disso, quando a quantidade de dados gerados aumenta, o $\mathcal{R}'_{\Upsilon}$ diminui, mostrando a escalabilidade do método proposto em relação à quantidade de dados sensoriados. Mais uma vez, o algoritmo se mostrou mais adequado utilizando as técnicas PCA e PCA-robusta, cujos resultados foram muito próximos. Com isso, é possível tomar as decisões D' relativas a $\mathcal{R}'_{\Upsilon}$.



(a) Redução $n/2$ - $\mathcal{V}'^{s}_{n/2}$

(b) Redução $\log_2 n$ - $\mathcal{V}'^{s}_{\log_2 n}$

FIGURA 10: $\mathcal{R}'_{\Upsilon}$ para redução no sensoriamento usando dados sintéticos sem ruído

5.2.1.2 Geração dos dados com ruído

Discute-se a seguir a redução no sensoriamento, na qual a geração dos dados é feita com a presença de ruído. Para isso, um ruído aleatório é inserido através do programa R durante a geração dos dados nas distribuição *normal*. Nesse caso, a geração dos dados foi feita a partir das mesmas cinco médias arbitrárias utilizadas na avaliação sem ruído $\mu = \{10, 30, 50, 70, 90\}$, escolhidas para simular o sensoriamento de cinco diferentes fenômenos. O tamanho dos dados foi variado em $n = \{256, 512,\ 1024, 2048\}$ e aplicou-se as reduções $n/2$ e $\log_2 n$, como no cenário anterior.

A primeira avaliação nesse cenário considera o $\mathcal{R}'_{\Phi}$. Valores observados para ambas as técnicas avaliadas são mostradas na tabela 2. Nessa avaliação, as amostragens baseadas em PCA, ICA e PCA-robusta tiveram resultados muito similares em todas as situações simuladas e esses resultados indicam que não existem diferenças significativas entre as variâncias dos conjuntos de dados original e reduzido para as três técnicas. As três técnicas apresentaram resultados satisfatórios em todos os casos, uma vez que a hipótese nula pode ser aceita. Com isso, para a utilização de dados com ruído, é possível tomar as decisões D' referentes à regra

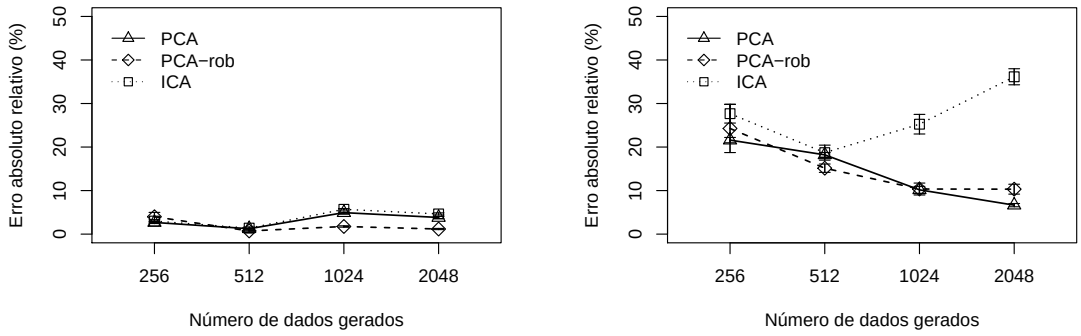
$\mathcal{R}'_{\Phi}$.

TABELA 2: Análise da variância utilizando dados sintéticos com ruído para redução no sensoriamento

Distribuição	$(n = 256)$		$(n = 512)$		$(n = 1024)$		$(n = 2048)$	
normal	$n/2$	$\log_2 n$	$n/2$	$\log_2 n$	$n/2$	$\log_2 n$	$n/2$	$\log_2 n$
PCA	0,38	0,55	0,29	0,48	0,19	0,61	0,09	0,51
ICA	0,38	0,56	0,26	0,46	0,19	0,60	0,08	0,50
PCA-rob	0,37	0,50	0,29	0,57	0,19	0,55	0,09	0,54

A segunda análise considera o $\mathcal{R}'_{\Upsilon}$ e os resultados podem ser observados na figura 11. Considerando a redução $n/2$ (figura 11(a)), mesmo com a presença de ruído nos dados, os resultados observados foram satisfatórios, sendo o maior valor encontrado para o $\mathcal{R}'_{\Upsilon}$ igual a aproximadamente 6%, quando utilizou-se a técnica ICA e gerou-se 1024 dados. Nesse caso, os resultados mais significativos foram observados quando empregou-se a técnica PCA-robusta e os resultados onde o algoritmo proposto se mostrou menos satisfatório foram observados com a técnica ICA. Como mencionado na avaliação sem ruído, a diferença de desempenho entre ICA e as demais técnicas pode ser explicada pelo fato de que a primeira componente independente pode não possuir a maior porcentagem de variância dos dados.

Considerando a redução $\log_2 n$ (figura 11(b)), os erros observados foram bastante superiores aos da redução $n/2$, principalmente quando a amostragem foi baseada em ICA, pelo mesmo motivo citado anteriormente. Para esta técnica, quando aumentou-se a quantidade de dados gerados, o $\mathcal{R}'_{\Upsilon}$ também aumentou em praticamente todos os casos. Entretanto, quando empregou-se as técnicas PCA e PCA-robusta, o algoritmo MuSA se mostrou mais eficiente e à medida que se aumentou a quantidade dos dados gerados, o $\mathcal{R}'_{\Upsilon}$ diminuiu em todos os casos, o que comprova a escalabilidade do algoritmo proposto em relação ao número de dados sensoriados, quando a amostragem é baseada nessas duas técnicas. Além disso, considerando que a variação dos dados originais é bastante significativa nesse caso, devido à introdução de ruído nesses dados, os resultados observados podem ser considerados satisfatórios, principalmente para maiores quantidades de dados gerados, sendo que, nesse caso, o $\mathcal{R}'_{\Upsilon}$ observado foi de aproximadamente 10%. Com isso, para a utilização de dados com ruído, as decisões D' referentes a $\mathcal{R}'_{\Upsilon}$ também podem ser tomadas.



(a) Redução $n/2$ - $\mathcal{V}'^{s}_{n/2}$

(b) Redução $\log_2 n$ - $\mathcal{V}'^{s}_{\log_2 n}$

FIGURA 11: $\mathcal{R}'_{\Upsilon}$ para redução no sensoriamento usando dados sintéticos sem ruído

5.2.2 Redução no nó líder

5.2.2.1 Geração dos dados sem ruído

O segundo cenário considera Ψ sendo efetuada no nó líder. Nesse caso, cada nó sensor i possui apenas um sensor que propaga suas leituras $\mathcal{V}^1_n$ do ambiente para um nó líder, que recebe as várias leituras dos diferentes nós, reduz os dados e os propaga até sorvedouro. Os dados $\mathcal{V}$ foram gerados com apenas uma média μ, utilizando a distribuição *normal* sem a presença de ruído, representando o sensoriamento de somente uma variável pelos diferentes sensores do agrupamento. Nesse caso, para cada nó do agrupamento, em cada simulação, uma sequência de dados foi gerada e todas as sequências têm a mesma média μ e desvio padrão σ. Os valores escolhidos foram $\mu = 50$ e $\sigma = 10\%$. Para as simulações, fixou-se o número de sensores do agrupamento em $s = 400$ e variou-se o tamanho dos dados gerados em $n = \{256, 512, 1024, 2048\}$, aplicando-se novamente as reduções $n/2$ e $\log_2 n$.

Assim como no caso da redução no sensoriamento, a primeira avaliação considera o $\mathcal{R}'_{\Phi}$ e os resultados são apresentados na tabela 3. Os resultados obtidos mostram que as amostragens baseadas em PCA, ICA e PCA-robusta não apresentam diferenças significativas entre as variâncias dos dados originais $\mathcal{V}$ e dos dados reduzidos $\mathcal{V}'$. Com isso, para a utilização de dados sem ruído, as decisões D' referentes a $\mathcal{R}'_{\Phi}$ podem ser tomadas.

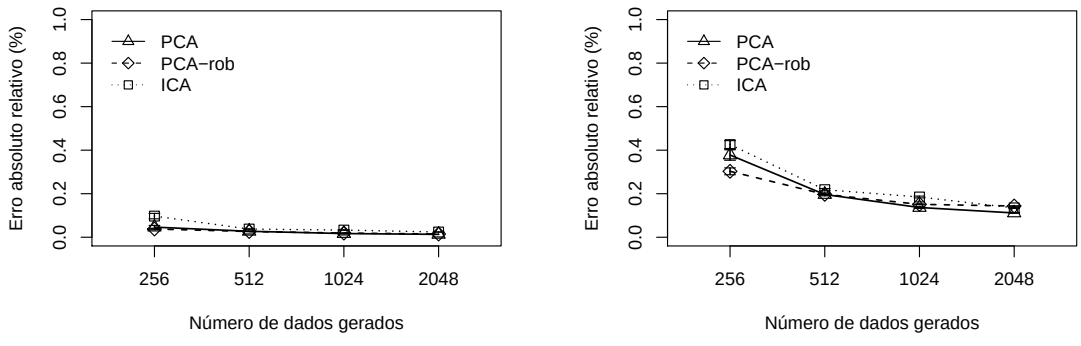
A segunda avaliação considera o $\mathcal{R}'_{\Upsilon}$. Na figura 12, apresenta-se os resultados obtidos para o algoritmo MuSA, com as técnicas PCA, ICA e PCA-robusta. Considerando a redução $n/2$, mostrada na figura 12(a), o algoritmo proposto se mostrou satisfatório com as três técnicas, sendo o $\mathcal{R}'_{\Upsilon}$ muito próximo de zero em todos os casos. Esse fato pode ser explicado porque todos os sensores no agrupamento monitoram o mesmo tipo de fenômeno. Dessa forma,

TABELA 3: Análise da variância utilizando dados sintéticos sem ruído com redução no nó líder

Distribuição	$(n = 256)$		$(n = 512)$		$(n = 1024)$		$(n = 2048)$	
normal	$n/2$	$\log_2 n$	$n/2$	$\log_2 n$	$n/2$	$\log_2 n$	$n/2$	$\log_2 n$
PCA	0,90	0,82	0,89	0,81	0,89	0,83	0,89	0,81
ICA	0,82	0,79	0,84	0,80	0,78	0,78	0,74	0,76
PCA-rob	0,90	0,81	0,89	0,86	0,89	0,84	0,89	0,82

uma maior quantidade de dados redundantes é encontrado e as técnicas conseguem melhor desempenho na seleção dos dados que irão compor o conjunto de dados reduzido. Nesse cenário, os resultados observados para as três técnicas foram novamente similares, e o maior $\mathcal{R}'_\Upsilon$ observado foi de aproximadamente $0,1\%$, com a amostragem baseada em ICA. Apesar da pequena diferença entre as técnicas, a amostragem baseada em PCA obteve melhor resultado na maioria dos casos.

Para as simulações utilizando a redução $\log_2 n$, mostrada na figura 12(b), os resultados observados comprovam a usabilidade do algoritmo MuSA com ambas as técnicas avaliadas. Nesse caso, assim como na redução $n/2$, os erros observados foram próximos de zero em todos os casos, sendo o maior valor encontrado para o $\mathcal{R}'_\Upsilon$ de aproximadamente $0,4\%$, quando a amostragem foi baseada em ICA e foram gerados 256 dados. Além disso, à medida que aumentou-se a quantidade de dados gerados, o $\mathcal{R}'_\Upsilon$ diminuiu em todos os casos, demonstrando uma escalabilidade do algoritmo MuSA em relação ao número de dados sensoriados, também para a redução no nó líder. Nesse caso, embora a diferença entre as técnicas seja muito pequena, a amostragem baseada em PCA obteve resultados melhores na maioria dos tamanhos de dados gerados. Com isso, para a utilização de dados sem ruído, as decisões D' referentes a $\mathcal{R}'_\Upsilon$ também podem ser tomadas.



(a) Redução $n/2$ - $\mathcal{V}'^{s}_{n/2}$

(b) Redução $\log_2 n$ - $\mathcal{V}'^{s}_{\log_2 n}$

FIGURA 12: $\mathcal{R}'_\Upsilon$ para redução no nó líder usando dados sintéticos sem ruído

5.2.2.2 Geração dos dados com ruído

A avaliação a seguir considera a redução no nó líder, na qual a geração dos dados é feita com a presença de ruído. Para isso, assim como na redução no sensoriamento, um ruído aleatório é inserido através do programa R durante a geração dos dados na distribuição *normal*. Como na avaliação sem ruído, os dados $\mathcal{V}_n^s$ foram gerados com apenas uma média μ, representando o sensoriamento de somente uma variável pelos diferentes sensores do agrupamento. Utilizou-se o mesmo valor $\mu = 50$, com um desvio padrão $\sigma = 10\%$. Para as simulações, fixou-se novamente o número de sensores do agrupamento em $s = 400$ e variou-se o tamanho dos dados gerados em $n = \{256, 512, 1024, 2048\}$, aplicando-se novamente as reduções $n/2$ e $\log_2 n$.

A primeira análise nesse cenário considera o $\mathcal{R}'_\Phi$ e os valores observados são apresentados na tabela 4. Resultados mostram que as amostragens baseadas nas técnicas PCA, ICA e PCA-robusta não apresentam diferenças significativas entre as variâncias dos dados originais e dos dados reduzidos. Nesse caso, diferentemente do ocorrido na avaliação da redução no sensoriamento com a presença de ruído, os resultados obtidos foram bastante significativos, com valores que se aproximam dos obtidos na avaliação sem ruído. Como mencionado anteriormente, isso ocorre pelo fato de que todos os sensores monitoram o mesmo fenômeno. Com isso, para a utilização de dados com ruído, é possível tomar as decisões D' referentes à regra $\mathcal{R}'_\Phi$.

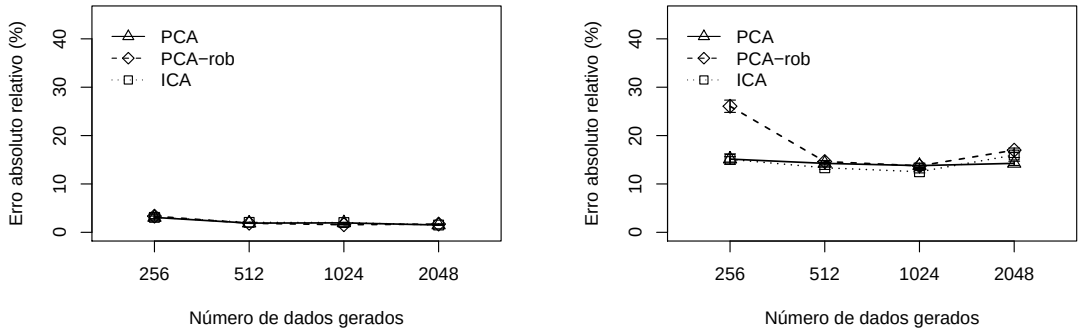
TABELA 4: Análise da variância utilizando dados sintéticos com ruído para redução no nó líder

Distribuição	$(n = 256)$		$(n = 512)$		$(n = 1024)$		$(n = 2048)$	
normal	$n/2$	$\log_2 n$	$n/2$	$\log_2 n$	$n/2$	$\log_2 n$	$n/2$	$\log_2 n$
PCA	0,64	0,62	0,61	0,64	0,60	0,65	0,57	0,64
ICA	0,64	0,61	0,61	0,62	0,59	0,66	0,57	0,63
PCA-rob	0,61	0,61	0,59	0,65	0,58	0,62	0,57	0,64

Considera-se agora a avaliação do $\mathcal{R}'_\Upsilon$, cujos valores são mostrados na figura 13. Considerando redução $n/2$ (figura 13(a)), os valores observados para as três técnicas foram novamente similares para todos os tamanhos de dados gerados. O algoritmo MuSA se mostrou bastante eficiente para a redução no nó líder, mesmo com a presença de ruído nos dados, o que eleva consideravelmente a variação nos mesmos. Nesse cenário, o maior $\mathcal{R}'_\Upsilon$ observado foi de aproximadamente 3.5%, quando utilizou-se a técnica PCA-robusta e gerou-se $\mathcal{V}_{256}^s$ dados.

Para as simulações com a redução $\log_2 n$ (figura 13(b)), os resultados observados com as três técnicas foram novamente similares, exceto quando gerou-se $\mathcal{V}_{256}^s$ dados e utilizou-se

a técnica PCA-robusta. Nesse caso, foi encontrado o maior $\mathcal{R}'_{\Upsilon}$, de aproximadamente 26%. Nos demais casos, os erros observados variaram entre 14% e 17%, aproximadamente, nas três técnicas. Mais uma vez, quando aumentou-se a quantidade de dados gerados, o $\mathcal{R}'_{\Upsilon}$ diminuiu ou se manteve próximo ao observado com uma menor quantidade de dados, o que ratifica a escalabilidade da solução proposta em termos da quantidade de dados sensoriados, também na redução no nó líder. Com isso, para a utilização de dados com ruído, é possível tomar as decisões D' referentes à regra $\mathcal{R}'_{\Upsilon}$.



(a) Redução $n/2$ - $\mathcal{V}'^{s}_{n/2}$

(b) Redução $\log_2 n$ - $\mathcal{V}'^{s}_{\log_2 n}$

FIGURA 13: $\mathcal{R}'_{\Upsilon}$ para redução no nó líder usando dados sintéticos com ruído

5.3 Dados reais

5.3.1 Considerações

Nesta seção, descreve-se a avaliação do algoritmo MuSA utilizando conjuntos de dados reais de uma aplicação de monitoramento da qualidade do ar. O conjunto de dados reais corresponde às médias de dados coletados durante dois dias, sendo cada valor a média de um fenômeno monitorado em um intervalo de quatro horas. Gerou-se dados pseudo-reais para cada sensor a partir de 12 médias μ reais disponibilizadas por Albuquerque (2007), considerando a distribuição *normal* multivariada. Os dados pseudo-reais representam a estimativa de valores coletados em cada intervalo de quatro horas. Com o objetivo de mostrar a variação desses dados pseudo-reais, a figura 14 apresenta o comportamento do algoritmo MuSA baseado em PCA, com esses valores de dados. Para essa observação, considerando apenas um fenômeno, foram gerados $\mathcal{V}^{1}_{256}$ dados com 6 médias μ reais coletadas durante um dia, em intervalos de quatro horas e aplicou-se as reduções $n/2$ e $\log_2 n$. Quando a redução utilizada é $n/2$, o algoritmo obtém uma variação próxima à observada com os dados originais. Quando a

redução aplicada é $\log_2 n$, a variação dos dados é perdida. Isso ocorre porque somente os valores intermediários são considerados na amostragem, ou seja, eventos inesperados não são percebidos, mostrando um indício de que para as aplicações de monitoramento da qualidade do ar, uma redução $\log_2 n$ não seja adequada. Entretanto, isso não implica que para outras aplicações essa redução também não seja adequada.

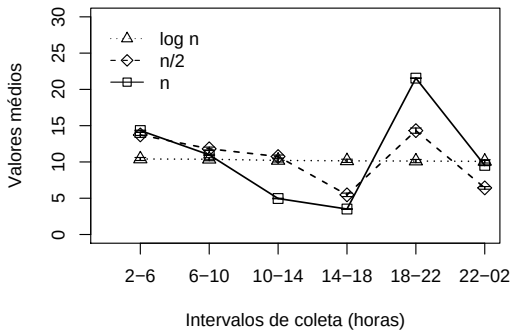


FIGURA 14: Comportamento do MuSA em termos da variação dos dados

Da mesma forma que a análise com dados sintéticos, considera-se a redução Ψ no sensoriamento e no nó líder. Mais uma vez, os resultados da avaliação de $\mathcal{R}'_{\Upsilon}$ são apresentados com um intervalo de confiança de 95%. Além disso, para cada uma das formas de redução, avaliou-se a representatividade dos dados com e sem a presença de ruído. As avaliações para cada um desses cenários são descritas a seguir.

5.3.2 Redução no sensoriamento

5.3.2.1 Geração dos dados sem ruído

A primeira avaliação considera a redução no momento do sensoriamento, na qual os dados são gerados sem a presença de ruído. Nesse caso, como na análise com dados sintéticos, simulou-se o monitoramento de 5 diferentes variáveis. Assim, para cada variável, os dados pseudo-reais foram gerados a partir de 12 médias reais μ, com $\sigma = 10\%$, usando a distribuição *normal*. A quantidade de dados gerados foi variada em $n = \{256, 512, 1024, 2048\}$ e aplicou-se as reduções $n/2$ e $\log_2 n$. É importante relembrar que os dados de entrada são representados por $\mathcal{V}_n^s$, onde n é a quantidade de dados sensoriados e s o número de sensores.

A primeira análise considera os resultados para o $\mathcal{R}'_{\Phi}$ e os resultados podem ser observados na tabela 5. Nesse caso, as amostragens baseadas em PCA, ICA e PCA-robusta não

apresentam diferenças significativas entre as variâncias de seus conjuntos de dados original e reduzido e seus resultados mostram um alto nível de significância. Com isso, utilizando dados pseudo-reais sem ruído, as decisões D' referentes a $\mathcal{R}'_{\Phi}$ podem ser tomadas.

TABELA 5: Análise da variância utilizando dados pseudo-reais sem ruído com redução no sensoriamento

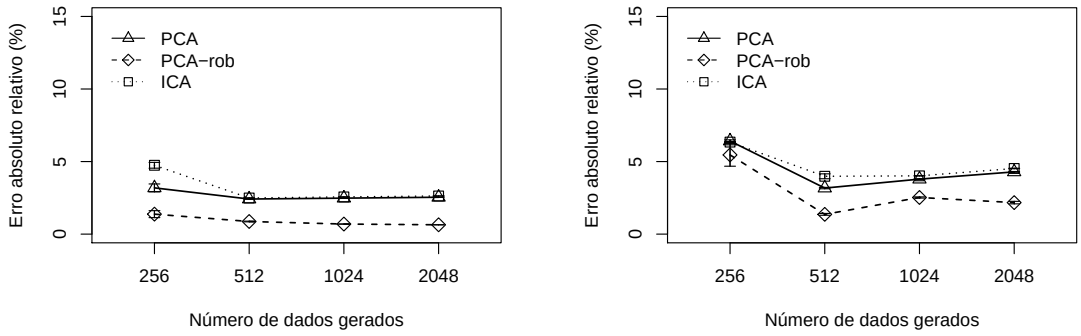
Distribuição	$(n = 256)$		$(n = 512)$		$(n = 1024)$		$(n = 2048)$	
normal	$n/2$	$\log_2 n$	$n/2$	$\log_2 n$	$n/2$	$\log_2 n$	$n/2$	$\log_2 n$
PCA	0,80	0,80	0,78	0,82	0,70	0,80	0,61	0,83
ICA	0,80	0,80	0,78	0,82	0,70	0,80	0,61	0,83
PCA-rob	0,77	0,79	0,65	0,77	0,57	0,79	0,49	0,76

A segunda análise nesse cenário considera o $\mathcal{R}'_{\Upsilon}$. Resultados de simulações são apresentados na figura 15. No que se refere à redução $n/2$, mostrada na figura 15(a), assim como no avaliação com dados sintéticos, o algoritmo MuSA teve resultados satisfatórios com ambas as técnicas utilizadas. Nesse cenário, o maior $\mathcal{R}'_{\Upsilon}$ foi observado com a amostragem baseada em ICA, no caso em que gerou-se $\mathcal{V}^5_{256}$ dados, sendo de aproximadamente 5%. Os melhores resultados foram obtidos pela técnica PCA-robusta, sendo sua superioridade observada para todos os tamanhos de dados gerados. Mais uma vez ficou comprovada a escalabilidade do algoritmo proposto em termos da quantidade de dados sensoriados, uma vez que, com ambas as técnicas utilizadas, quando aumentou-se essa quantidade de dados $\mathcal{V}$, o $\mathcal{R}'_{\Upsilon}$ diminuiu ou se manteve praticamente o mesmo.

Considerando as simulações com a redução $\log_2 n$, como pode ser visto na figura 15(b), a superioridade da amostragem baseada em PCA-robusta ficou comprovada nesse caso. Merece destaque nesse cenário o fato de que os $\mathcal{R}'_{\Upsilon}$ observados com a redução $\log_2 n$ foram baixos, sendo próximos aos observados com a redução $n/2$, ratificando a viabilidade do uso do algoritmo MuSA para efetuar a redução considerando o monitoramento de diferentes fenômenos. Nesse caso, o maior $\mathcal{R}'_{\Upsilon}$ encontrado foi de aproximadamente $6,5\%$, utilizando as técnicas PCA e ICA, gerando-se $\mathcal{V}^5_{256}$ dados. Com isso, para a utilização de dados pseudo-reais sem ruído, as decisões D' referentes a $\mathcal{R}'_{\Upsilon}$ podem ser tomadas.

5.3.2.2 Geração dos dados com ruído

A avaliação a seguir retrata a redução no sensoriamento, na qual os dados pseudo-reais são gerados com a presença de ruído. Da mesma forma que na avaliação com dados sintéticos, foi introduzido um ruído aleatório nos dados gerados com a distribuição *normal*, através do programa R. Novamente, simulou-se o monitoramento de 5 diferentes variáveis e, para cada



(a) Redução $n/2$ - $\mathcal{V}'^{s}_{n/2}$ (b) Redução $\log_2 n$ - $\mathcal{V}'^{s}_{\log_2 n}$

FIGURA 15: $\mathcal{R}'_{\Upsilon}$ para redução no sensoriamento usando dados pseudo-reais sem ruído

variável, os dados pseudo-reais foram gerados a partir de 12 médias reais μ, com $\sigma = 10\%$. A quantidade de dados gerados foi variada em $n = \{256, 512, 1024, 2048\}$ e aplicou-se as reduções $n/2$ e $\log_2 n$.

Avaliou-se primeiramente nesse cenário o $\mathcal{R}'_{\Phi}$ e os resultados obtidos são apresentados na tabela 6. Mais uma vez, os resultados das técnicas avaliadas foram similares com todos os tamanhos de dados gerados e reduções aplicadas. Nesse caso, o nível de significância dos resultados obtidos é menor, se comparado com os resultados da avaliação sem ruído, o que pode ser explicado pelo significativo aumento da variação dos dados provocada pela introdução do ruído nos mesmos. Entretanto, os resultados ainda podem ser considerados satisfatórios, pois todos os valores estão acima de $0,05$. Com isso, para a utilização de dados pseudo-reais com ruído, as decisões D' referentes a $\mathcal{R}'_{\Phi}$ podem ser tomadas.

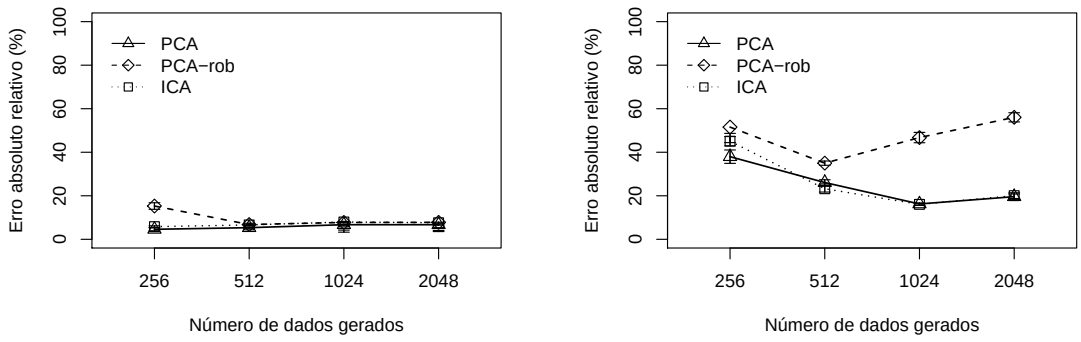
TABELA 6: Análise da variância utilizando dados pseudo-reais com ruído para redução no sensoriamento

Distribuição	$(n = 256)$		$(n = 512)$		$(n = 1024)$		$(n = 2048)$	
normal	$n/2$	$\log_2 n$	$n/2$	$\log_2 n$	$n/2$	$\log_2 n$	$n/2$	$\log_2 n$
PCA	0,36	0,52	0,12	0,55	0,09	0,58	0,07	0,53
ICA	0,36	0,54	0,12	0,55	0,08	0,58	0,07	0,51
PCA-rob	0,35	0,54	0,12	0,55	0,09	0,54	0,07	0,55

A segunda análise nesse cenário considera o $\mathcal{R}'_{\Upsilon}$, cujos resultados são mostrados na figura 16. Considerando os resultados da avaliação com a redução $n/2$, apresentados na figura 16(a), o melhor resultado foi observado com a amostragem baseada em PCA. Os resultados das três técnicas foram similares, exceto quando gerou-se $\mathcal{V}^5_{256}$ dados, cujo $\mathcal{R}'_{\Upsilon}$ observado com a técnica PCA-robusta foi o maior entre todos os casos avaliados, sendo de aproximadamente 15%. É importante destacar que, embora a variação dos dados seja ainda mais

considerável devido ao ruído, os resultados obtidos foram satisfatórios, reforçando a viabilidade do uso do método proposto quando da aplicação da redução $n/2$.

No que se refere à redução $\log_2 n$, mostrada na figura 16(b), o algoritmo novamente se mostrou menos adequado quando utilizou a técnica PCA-robusta. Com essa técnica foi observado um $\mathcal{R}'_\Upsilon$ de aproximadamente 57% - o maior entre todos os valores observados. Utilizando as técnicas PCA e ICA, os resultados foram similares, havendo uma pequena superioridade de PCA na maioria dos casos. Nesse cenário, embora os erros observados sejam significativos, os mesmos ainda podem ser considerados satisfatórios, principalmente quando utilizou-se as técnicas PCA e ICA, uma vez que a variação dos dados foi muito significativa. Com isso, para a utilização de dados pseudo reais com ruído, as decisões D' referentes a $\mathcal{R}'_\Upsilon$ podem ser tomadas, especialmente quando da aplicação da redução $n/2$.



(a) Redução $n/2$ - $\mathcal{V}'^{s}_{n/2}$

(b) Redução $\log_2 n$ - $\mathcal{V}'^{s}_{\log_2 n}$

FIGURA 16: $\mathcal{R}'_\Upsilon$ para redução no sensoriamento usando dados pseudo-reais com ruído

5.3.3 Redução no nó líder

5.3.3.1 Geração dos dados sem ruído

A avaliação a seguir diz respeito à redução no nó líder, na qual a geração dos dados se dá sem a presença de ruído. Os dados pseudo-reais foram gerados representando o sensoriamento de uma única variável por diferentes nós sensores em um agrupamento. Nesse caso, o agrupamento é composto por 400 nós e para cada nó, o dado pseudo-real considera 12 médias reais μ, com $\sigma = 10\%$ e usando a distribuição *normal*. Assim como nas avaliações anteriores, variou-se o tamanho dos dados gerados em $n = \{256, 512, 1024, 2048\}$ e aplicou-se as reduções $n/2$ e $\log_2 n$.

Avaliou-se primeiramente os resultados para o $\mathcal{R}'_{\Phi}$, cujos valores são mostrados na tabela 7. Assim como na avaliação com dados sintéticos, os resultados mostram que não existem diferenças significativas entre as variâncias de seus conjuntos de dados original e reduzido considerando as três técnicas empregadas e, em todos os casos, os resultados apresentarem alto nível de significância. Com isso, para a utilização de dados pseudo-reais sem ruído, as decisões D' referentes a $\mathcal{R}'_{\Phi}$ podem ser tomadas.

TABELA 7: Análise da variância utilizando dados pseudo-reais sem ruído com redução no nó líder

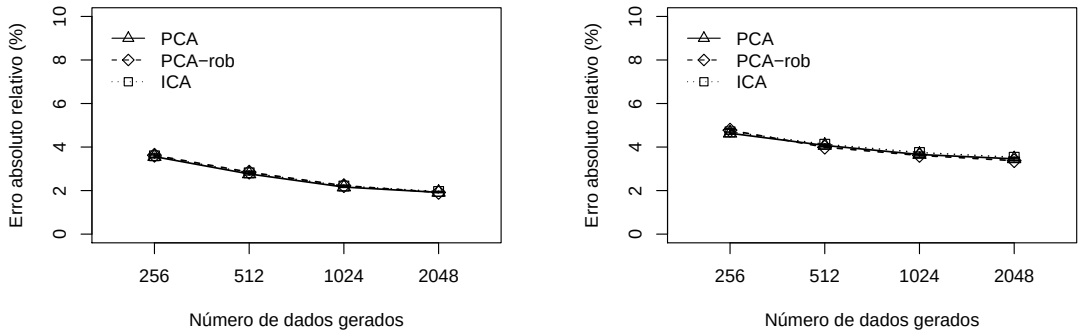
Distribuição	$(n = 256)$		$(n = 512)$		$(n = 1024)$		$(n = 2048)$	
normal	$n/2$	$\log_2 n$	$n/2$	$\log_2 n$	$n/2$	$\log_2 n$	$n/2$	$\log_2 n$
PCA	0,78	0,77	0,75	0,79	0,69	0,76	0,61	0,80
ICA	0,78	0,76	0,71	0,79	0,67	0,75	0,59	0,80
PCA-rob	0,76	0,74	0,68	0,75	0,64	0,73	0,56	0,77

A segunda análise nesse cenário considera o $\mathcal{R}'_{\Upsilon}$ e os resultados observados são apresentados na figura 17. Considerando a redução $n/2$, mostrada na figura 17(a), as técnicas PCA, ICA e PCA-robusta apresentaram resultados praticamente idênticos para todos os tamanhos de dados gerados, e os valores observados para o $\mathcal{R}'_{\Upsilon}$ mostram a viabilidade do uso do algoritmo MuSA nesse cenário. Nesse caso, o maior $\mathcal{R}'_{\Upsilon}$ encontrado foi de aproximadamente 4%, o que pode ser considerado como satisfatório, especialmente pelo fato de que a variação nos dados originais foi considerável.

Os resultados observados na avaliação da redução $\log_2 n$, mostrados na figura 17(b), foram similares aos observados com a redução $n/2$, o que também reforça a viabilidade do uso do algoritmo proposto. Nesse cenário, o maior $\mathcal{R}'_{\Upsilon}$ observado foi de aproximadamente 5%, valor muito baixo se considerarmos o tamanho da amostra dos dados que seria enviada ao sorvedouro nesse caso. Além disso, quando aumentou-se a quantidade de dados gerados, o $\mathcal{R}'_{\Upsilon}$ diminuiu em todos os casos, mostrando mais uma vez a escalabilidade do MuSA em termos da quantidade de dados sensoriados. Com isso, para a utilização de dados pseudo-reais sem ruído, as decisões D' referentes a $\mathcal{R}'_{\Upsilon}$ também podem ser tomadas.

5.3.3.2 Geração dos dados com ruído

Discute-se a seguir a avaliação da redução no nó líder, com os dados sendo gerados com através do programa R, com a introdução de um ruído aleatório nesses dados. Assim como na avaliação com os dados sem ruído, gerou-se dados representando o sensoriamento de uma única variável por diferentes nós sensores em um agrupamento. O agrupamento é composto



(a) Redução $n/2$ - $\mathcal{V}^{\prime s}_{n/2}$

(b) Redução $\log_2 n$ - $\mathcal{V}^{\prime s}_{\log_2 n}$

FIGURA 17: $\mathcal{R}'_{\Upsilon}$ para redução no nó líder usando dados pseudo-reais sem ruído

por 400 nós e para cada nó, o dado pseudo-real $\mathcal{V}$ considera 12 médias reais μ, com $\sigma = 10\%$ e usando a distribuição *normal*. Mais uma vez, variou-se o tamanho dos dados gerados em $n = \{256, 512, 1024, 2048\}$ e aplicou-se as reduções $n/2$ e $\log_2 n$.

A primeira análise nesse cenário é feita sobre o $\mathcal{R}'_{\Phi}$, que tem seus resultados apresentados na tabela 8. Da mesma forma que nos cenários anteriores, os resultados mostraram que não existem diferenças significativas entre as variâncias dos conjuntos de dados original e reduzido em todos os casos simulados. Com isso, para a utilização de dados pseudo-reais com ruído, as decisões D' referentes a $\mathcal{R}'_{\Phi}$ podem ser tomadas.

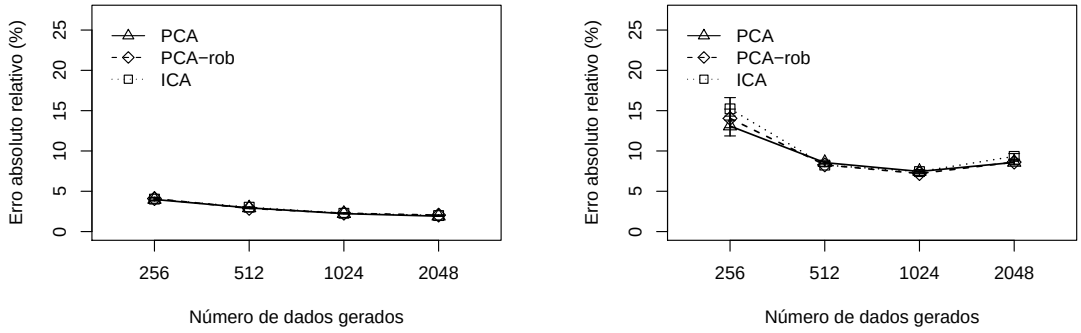
TABELA 8: Análise da variância utilizando dados pseudo-reais com ruído para redução no nó líder

Distribuição	$(n = 256)$		$(n = 512)$		$(n = 1024)$		$(n = 2048)$	
normal	$n/2$	$\log_2 n$	$n/2$	$\log_2 n$	$n/2$	$\log_2 n$	$n/2$	$\log_2 n$
PCA	0,12	0,55	0,13	0,52	0,45	0,66	0,59	0,62
ICA	0,12	0,52	0,13	0,52	0,45	0,66	0,59	0,63
PCA	0,11	0,63	0,15	0,57	0,50	0,63	0,56	0,64

A segunda análise, cujos resultados são mostrados na figura 18, é feita sobre o $\mathcal{R}'_{\Upsilon}$. Em relação à redução $n/2$, mostrada na figura 18(a), os resultados observados foram similares aos da avaliação com dados sem ruído. Nesse cenário, as técnicas PCA, ICA e PCA-robusta apresentaram resultados bastante satisfatórios e praticamente idênticos em todos os casos, sendo o maior $\mathcal{R}'_{\Upsilon}$ encontrado de aproximadamente 4%. Ficou também comprovada a escalabilidade da solução em termos da quantidade de dados sensoriados, visto que quando aumentou-se essa quantidade de dados, o $\mathcal{R}'_{\Upsilon}$ diminuiu em todos os casos.

Considerando a redução $\log_2 n$, que tem seus resultados apresentados na figura 18(b), os

resultados obtidos com as três técnicas analisadas foram novamente similares, havendo uma pequena superioridade da amostragem baseada em PCA. Os resultados mostraram ainda que o algoritmo MuSA é bastante adequado para redução nó líder, mesmo quando a variação nos dados é muito grande, como nesse caso. O maior valor observado para o $\mathcal{R}'_{\Upsilon}$ foi de aproximadamente 16%, utilizando-se a técnica ICA, com $\mathcal{V}^{s}_{256}$ dados gerados. Com isso, para a utilização de dados pseudo-reais com ruído, as decisões D' referentes a $\mathcal{R}'_{\Upsilon}$ também podem ser tomadas.



(a) Redução $n/2$ - $\mathcal{V}'^{s}_{n/2}$

(b) Redução $\log_2 n$ - $\mathcal{V}'^{s}_{\log_2 n}$

FIGURA 18: $\mathcal{R}'_{\Upsilon}$ para redução no nó líder usando dados pseudo-reais com ruído

5.4 Conclusões parciais

Considerando os resultados apresentados, o algoritmo MuSA se mostrou satisfatório em todos os cenários avaliados, tanto para a análise da variância dos dados $\mathcal{R}'_{\Phi}$, como do erro absoluto relativo $\mathcal{R}'_{\Upsilon}$, inclusive quando a variação dos dados foi considerável, como na avaliação dos dados com ruído. Nesse caso, pode-se dizer que a partir do conjunto de regras $\mathcal{R}'$, é possível tomar um conjunto de decisões D', que são correspondentes ao conjunto de decisões D que seriam tomadas caso todos os dados sensoriados fossem transmitidos. Comparando as técnicas PCA, ICA e PCA-robusta, as amostragens baseadas em PCA e PCA-robusta se mostraram mais eficientes que a baseada em ICA, principalmente na redução no sensoriamento. Nesse caso, a técnica PCA merece destaque, uma vez que, mesmo quando considerou-se a geração dos dados com ruído, os resultados obtidos com essa técnica estiveram sempre entre os melhores. Além disso, a técnica proposta se mostrou escalável em termos da quantidade de dados sensoriados, uma vez que mesmo aumentando a quantidade de dados gerados, os erros observados mantiveram-se aceitáveis. Para a redução no nó líder, os resultados observados foram similares na maioria dos cenários. Nesse caso, como na redução no sensoriamento, a

amostragem baseada em PCA também merece destaque, devido aos melhores resultados em grande parte das simulações realizadas.

É importante enfatizar que, apesar dos resultados da amostragem baseada em ICA ter desempenho inferior na maioria dos cenários, considerando as características descritas na seção 2.2.2, é possível que a mesma obtenha resultados melhores caso diferentes fenômenos sejam monitorados e os mesmos sejam não gaussianos. Outra possibilidade de melhorar o desempenho da técnica é efetuar a ordenação das componentes independentes, não fazendo, entretanto, parte do escopo deste trabalho. Outro ponto importante é que, apesar da técnica PCA-robusta ter sido proposta para melhorar o desempenho de PCA quando da presença de valores discrepantes ou atípicos nos dados, em alguns casos a técnica foi menos eficiente que a PCA tradicional. Isso pode ser explicado pelo fato de que o cenário no qual as técnicas foram utilizadas é diferente do abordado na maioria das vezes, que é a redução de dimensionalidade.

6 COMPORTAMENTO DA REDE

Este capítulo tem por objetivo descrever a avaliação do comportamento da rede ao se efetuar a redução dos dados multivariados através do algoritmo MuSA. Uma vez que a comunicação é a tarefa que mais consome energia nas redes de sensores sem fio, reduzindo a quantidade de dados transmitidos, naturalmente reduz-se a quantidade de energia consumida. Para essa avaliação, as simulações são feitas utilizando apenas o algoritmo baseado em PCA, pois para a avaliação do comportamento da rede, as diferenças entre o processamento realizado pelas demais técnicas não afeta de forma significativa o desempenho do algoritmo, uma vez que o modelo de dissipação de energia utilizado não leva em consideração o consumo de energia do processamento. No simulador utilizado, o modelo de dissipação de energia é linear, com valores pré-definidos para transmissão e recepção. Neste trabalho, o comportamento da rede é avaliado para mostrar os benefícios da redução de dados em termos do consumo de energia e do atraso para entrega dos dados ao sorvedouro. Da mesma forma que foi feito na avaliação da representatividade dos dados, discute-se a respeito da aplicação da redução Ψ no momento do sensoriamento e no nó líder.

6.1 Redução no momento do sensoriamento

A primeira avaliação do comportamento da rede se refere à redução no sensoriamento. A avaliação do algoritmo nesse cenário é baseada nas seguintes considerações:

- **Simulação**: a avaliação do comportamento da rede se dá através de simulações utilizando o NS-2 (*Network Simulator 2*) versão 2.33*. A simulação foi executada com 33 topologias aleatórias e os resultados são apresentados com intervalo de confiança de 95%.
- **Topologia da rede**: utilizou-se uma rede plana com um algoritmo de roteamento baseado em árvore de menor caminho e todos os nós possuem a mesma configuração de *hardware*. A densidade da rede é mantida constante e os nós fontes são aleatoriamente distribuídos na região sensoriada. Afim de avaliar somente o desempenho na redução dos dados, as árvores são construídas somente uma vez antes que o tráfego se inicie.

*http://nsnam.isi.edu/nsnam/index.php/Main_Page

- **Tipo de dados**: considerou-se como dados de entrada $\mathcal{V}_n^s$, onde $n > 0$ é a quantidade de dados sensoriados e $s > 1$ é o número de sensores em cada nó da rede. Os dados sensoriados por cada sensor seguem uma distribuição *normal*, variando entre [0:1], com $\mu = 0.5$ e $\sigma = 0.1$. Por convenção, o tamanho dos pacotes de dados suportados pelo nó sensor foi definido como 20 bytes. Quando a quantidade de dados gerados $\mathcal{V}_n^s$ é maior que o tamanho do pacote, eles são fragmentados e reconstruídos na recepção pelo sorvedouro.

- **Parâmetros avaliados**: variou-se o tamanho dos dados em $n = \{256, 512, 1024, 2048\}$, com um número fixo de sensores $s = 5$. O número de nós na rede foi variado em $\{128, 256, 512, 1024\}$ e ainda variou-se o número de nós fonte em $\{1, 5, 10, 20\}$. Para avaliar o comportamento da rede utilizando o algoritmo MuSA, foram aplicadas as reduções $n/2$ e $\log_2 n$. Alguns parâmetros importantes utilizados nas simulações são apresentados na tabela 9.

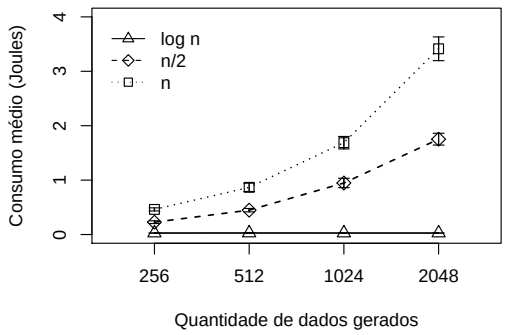
TABELA 9: Parâmetros de simulação

Parametro	**Valores**
Tamanho da rede	varia com a densidade
Tamanho da fila	varia com o tamanho do dado
Localização do sorvedouro	coordenadas (0, 0)
Localização do nó fonte	aleatória
Alcance do rádio (m)	50
Largura de banda (kbps)	250
Tempo de simulação (s)	1100
Início do tráfego (s)	500
Fim do tráfego (s)	600
Energia inicial (J)	100

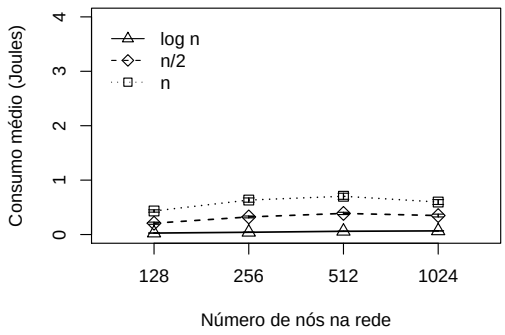
O tamanho da rede varia para manter a densidade da rede constante em $8,48$ e é obtido através de $rede_t = \sqrt{\pi\, a_r^2\, |S|/8,4791}$, onde a_r é o alcance do rádio e S o número de nós sensores. O tamanho da fila suportada por cada nó da rede varia com o tamanho do dado para que não haja descarte de pacotes, ou seja, cada nó suporta o tamanho do dado utilizado pela aplicação. O alcance do rádio e a largura de banda consideram a especificação do sensor MicaZ[†]. O tempo de simulação foi fixado em 1100 s, onde os 500 s iniciais são utilizados para montagem e configuração da rede e da estrutura de roteamento, os 500 s finais são utilizados para permitir que os pacotes restantes na rede sejam transmitidos. Com isso, o tráfego real de dados na rede dura 100 s. Além disso, a energia inicial utilizada foi 100 J, para que os nós da rede nunca tivessem suas reservas de energia esgotadas.

[†]http://www.xbow.com/

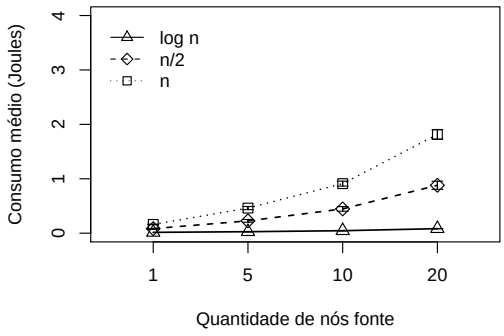
A primeira avaliação da redução no sensoriamento se refere ao consumo médio de energia da rede. Para essa avaliação, simulou-se o algoritmo MuSA em três diferentes situações: variando-se o tamanho dos dados sensoriados, variando a quantidade de nós na rede e também a quantidade de nós fonte, ou seja, nós que geram dados na rede. Em todas as avaliações, os resultados mostram o envio dos $\mathcal{V}_n^s$ dados gerados e as situações em que aplicou-se as reduções $n/2$ e $\log_2 n$, resultando em $\mathcal{V}'^s_{n/2}$ e $\mathcal{V}'^s_{\log_2 n}$, respectivamente. Os resultados das simulações são apresentados na figura 19.



(a) Variando o tamanho dos dados



(b) Variando número de nós na rede



(c) Variando número de nós fonte

FIGURA 19: Avaliação do consumo de energia médio na rede ao reduzir os dados

A primeira análise, mostrada na figura 19(a), refere-se à simulação feita variando-se o tamanho dos dados. Nesse caso, variou-se esse tamanho em $n = \{256, 512, 1024, 2048\}$, utilizando um número fixo de sensores por nó $s = 5$. O número de nós fonte foi fixado em 5 e a quantidade de nós da rede em 128. É possível observar que o consumo de energia aumenta significativamente quando se aumenta o tamanho dos dados gerados, o que ocorre porque o tráfego inserido na rede também aumenta. Entretanto, com a redução $\log_2 n$ essa variação não é observada, uma vez que a quantidade de tráfego na rede é muito pequeno.

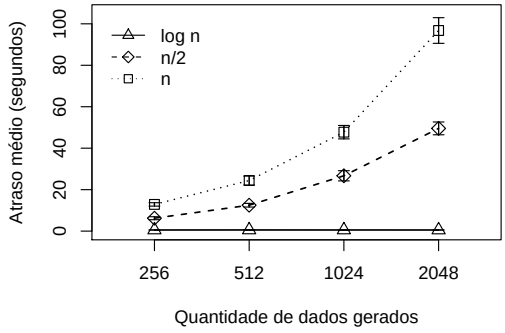
Por exemplo, para $\mathcal{V}^5_{256}$, somente 40 elementos são enviados em dois pacotes e para $\mathcal{V}^5_{2048}$, somente 50 elementos são enviados em três pacotes.

A segunda análise é realizada variando-se o número de nós na rede e os resultados são apresentados na figura 19(b). Para essa análise, variou-se o número de nós em $\{128, 256, 512, 1024\}$ novamente com um número fixo de sensores $s = 5$. Fixou-se ainda a quantidade de dados gerados em $n = 256$ e o número de nós fonte em 5. Percebe-se, nesse caso, que o consumo de energia permanece praticamente constante à medida que a quantidade de nós na rede é aumentada, uma vez que o tráfego na rede também é praticamente o mesmo. No entanto, um pequeno aumento no consumo de energia pode ser observado ao se enviar os n dados, quando aumenta-se a quantidade de nós na rede. Isso ocorre porque um número maior de nós são requisitados para encaminhar os pacotes ao sorvedouro.

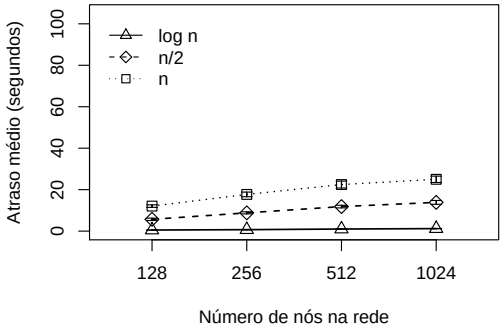
Por fim, avaliou-se nesse cenário, o consumo de energia variando-se o número de nós fonte, mostrado na figura 19(c). Para isso, variou-se o número de nós fonte em $\{1, 5, 10, 20\}$, com um número fixo de sensores $s = 5$ por nó da rede. Fixou-se ainda a quantidade de dados gerados em $n = 256$ e o número de nós em 128. Nesse caso, o consumo de energia aumenta consideravelmente quando aumenta-se a quantidade de nós fonte. Isso ocorre porque a quantidade de tráfego na rede também aumenta consideravelmente quando esse número de fontes aumenta. Mais uma vez, é possível notar que o consumo de energia diminui à medida que se reduz a quantidade de dados transmitidos. Além disso, novamente percebe-se que quando se aplica a redução $\log_2 n$, praticamente não existe variação do consumo de energia, uma vez que a quantidade de dados enviados na rede se mantém muito pequena.

A segunda avaliação da redução no sensoriamento se refere ao atraso médio na entrega dos dados ao sorvedouro. Para essa avaliação, foram considerados os mesmos cenários descritos na avaliação do consumo de energia. Os resultados das simulações são apresentados na figura 20.

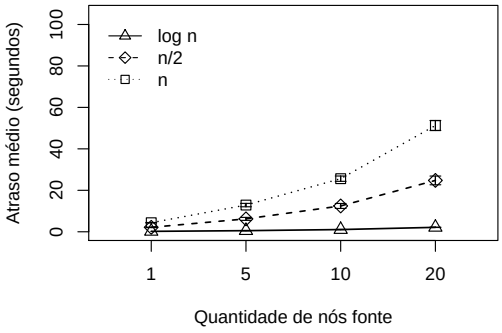
A primeira análise, mostrada na figura 20(a), se refere à simulação feita variando-se o tamanho dos dados. Nesse caso, variou-se esse tamanho em $n = \{256, 512, 1024, 2048\}$, utilizando um número fixo de sensores por nó $s = 5$. O número de nós fonte foi fixado em 5 e a quantidade de nós da rede em 128. Assim como ocorreu com o consumo de energia, quando diminui-se a quantidade de dados enviados, o atraso na entrega dos dados ao sorvedouro também diminui consideravelmente. A mesma variação do atraso também pode ser observada quando o tamanho dos dados gerados aumenta, pelo fato de que uma quantidade maior de pacotes serão enviados. Mais uma vez, com a redução $\log_2 n$ essa variação não é observada, uma vez que a quantidade de tráfego na rede é muito pequena.



(a) Variando o tamanho dos dados



(b) Variando número de nós na rede



(c) Variando número de nós fonte

FIGURA 20: Avaliação do atraso médio na rede ao reduzir os dados

A segunda análise é realizada variando-se o número de nós na rede e os resultados são apresentados na figura 20(b). Para essa análise, variou-se o número de nós em $\{128, 256, 512, 1024\}$, novamente com um número fixo de sensores $s = 5$. Fixou-se ainda a quantidade de dados gerados em $n = 256$ e o número de nós fonte em 5. Nesse caso também se percebe a mesma relação observada para o consumo de energia, ou seja, que o atraso varia muito pouco à medida que a quantidade de nós na rede é aumentada, uma vez que o tráfego na rede também é praticamente o mesmo. Um pequeno aumento pode ser observado nos casos em que se envia os n dados sensoriados ou em que se aplica a redução $n/2$, quando aumenta-se a quantidade de nós na rede.

Avaliou-se ainda nesse cenário, o atraso variando-se o número de nós fonte, como mostrado na figura 20(c). Mais uma vez, variou-se o número de nós fonte em $\{1, 5, 10, 20\}$, com um número fixo de sensores $s = 5$ por nó da rede. Fixou-se ainda a quantidade de dados gerados em $n = 256$ e o número de nós em 128. Novamente os resultados mostram a mesma relação observada na avaliação do consumo de energia. Nesse caso, o atraso aumenta

consideravelmente quando aumenta-se a quantidade de nós fonte, pelo mesmo motivo citado no cenário anterior. Mais uma vez, é possível notar também que o atraso diminui à medida que se reduz a quantidade de dados transmitidos. Além disso, novamente percebe-se que quando se aplica a redução $\log_2 n$, praticamente não existe variação do atraso, uma vez que a quantidade de pacotes transmitidos na rede se mantém muito pequena.

6.2 Redução no nó líder

A segunda avaliação do comportamento da rede ao se efetuar a redução de dados se refere à redução no nó líder. A avaliação do algoritmo nesse cenário é baseada nas seguintes considerações:

- **Simulação**: como na avaliação da redução no sensoriamento, a avaliação do comportamento da rede se dá através de simulações utilizando o NS-2 (*Network Simulator 2*) versão 2.33. A simulação foi executada com 33 topologias aleatórias e os resultados são apresentados com intervalo de confiança de 95%.

- **Topologia da rede**: utilizou-se uma rede hierárquica, com o roteamento baseado em árvore dentro e fora do agrupamento. A densidade da rede é mantida constante e todos os nós possuem a mesma configuração de *hardware*. As árvores de roteamento são construídas somente uma vez antes que o tráfego se inicie.

- **Tipo de dados**: considerou-se novamente como dados de entrada $\mathcal{V}_n^s$, onde n é a quantidade de dados sensoriados e s é o número de nós em cada agrupamento da rede. Utilizou-se, nesse caso, os mesmos parâmetros considerados na redução no sensoriamento. Quando os dados gerados $\mathcal{V}_n^s$ são maiores que o tamanho do pacote, eles são fragmentados e reconstruídos na recepção pelo nó líder e, posteriormente, pelo sorvedouro.

- **Parâmetros avaliados**: variou-se o tamanho dos dados suportados pelo líder em $n = \{256,\ 512, 1024, 2048\}$, o número de nós por agrupamento na rede em $s = \{128, 256, 512, 1024\}$ e o número de agrupamentos em $\{2, 4, 6, 8\}$. Nesse caso, todos os nós presentes nos agrupamentos enviam seus dados monitorados para o líder. Para avaliar o comportamento da rede utilizando o algoritmo MuSA, foram aplicadas as reduções $n/2$ e $\log_2 n$. Parâmetros importantes utilizados nas simulações são apresentados na tabela 10.

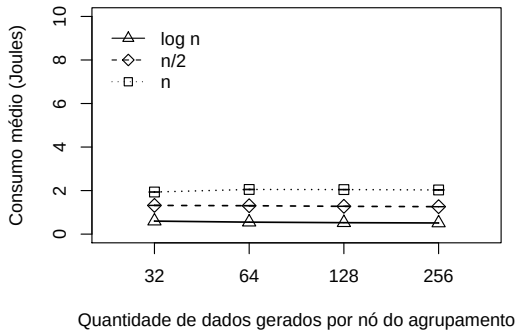
TABELA 10: Parâmetros de simulação

Parametro	**Valores**
Tamanho da rede	varia com a densidade
Tamanho da fila	varia com o tamanho do dado
Localização do sorvedouro	coordenadas (0, 0)
Alcance do rádio (m)	50
Largura de banda (kbps)	250
Tempo de simulação (s)	5000
Início do tráfego (s)	1000
Fim do tráfego (s)	4000
Energia inicial (J)	1000

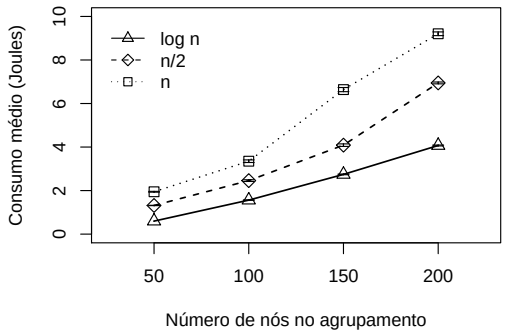
O tamanho da rede varia de acordo com a densidade e é obtido através de $rede_t = \sqrt{\pi\, a_r^2\, |S|/8,4791}$, onde a_r é o alcance do rádio e S o número de nós sensores. O tamanho da fila suportada pelo nó líder varia com o tamanho dos dados, para que não haja descarte de pacotes. O alcance do rádio e a largura de banda seguem a especificação do MicaZ. O tempo de simulação foi fixado em 5000 s, onde os 1000 s iniciais são utilizados para montagem e configuração da rede e da estrutura de roteamento, os 1000 s finais são utilizados para permitir que os pacotes restantes na rede sejam transmitidos. Com isso, o tráfego real de dados na rede dura 3000 s. Além disso, a energia inicial utilizada foi 100 J, para que os nós da rede nunca tivessem suas reservas de energia esgotadas.

A primeira avaliação da redução no nó líder se refere ao consumo médio de energia da rede. Para essa avaliação, simulou-se o algoritmo MuSA em três diferentes situações: variando-se o tamanho dos dados enviados pelos nós do agrupamento ao nó líder, variando a quantidade de nós por agrupamento e também a quantidade de agrupamentos na rede. Nessa avaliação, cada nó da rede possui apenas um sensor e todos monitoram o mesmo fenômeno. Como na redução no sensoriamento, em todos as avaliações, os resultados mostram o envio dos $\mathcal{V}_n^s$ dados gerados e as situações em que aplicou-se as reduções $n/2$ e $\log_2 n$, resultando em $\mathcal{V}_{n/2}'^s$ e $\mathcal{V}_{\log_2 n}'^s$, respectivamente. Os resultados das simulações são apresentados na figura 21.

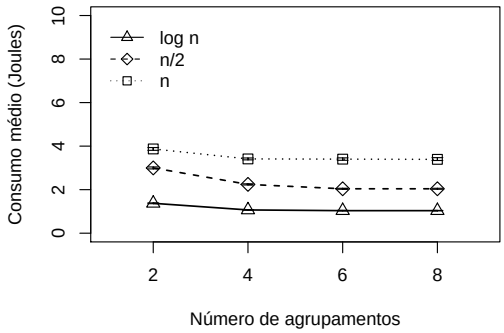
A primeira análise, mostrada na figura 21(a), se refere à simulação feita variando-se o tamanho dos dados enviados pelos nós do agrupamento ao nó líder. Nesse caso, variou-se esse tamanho em $n = \{32, 64, 128, 256\}$, utilizando um número fixo de nós por agrupamento em $s = 50$ e a quantidade de agrupamentos na rede em 2. O tamanho dos dados suportados pelo líder é dado pela multiplicação do número de nós do agrupamento pelo tamanho dos dados enviados por cada nó e, com isso, variou em $n = \{1600, 3200, 6400, 12800\}$. Nessa avaliação, quando aumentou-se o tamanho dos dados, a energia consumida se manteve praticamente constante, pois tem-se somente 2 agrupamentos, gerando um baixo tráfego na rede. Mais uma vez, quanto menor a quantidade de dados enviados pelo nó líder, ou seja, quanto maior



(a) Variando o tamanho dos dados enviados pelos nós do agrupamento ao nó líder



(b) Variando número de nós por agrupamento



(c) Variando número de agrupamentos

FIGURA 21: Avaliação do consumo de energia médio na rede ao reduzir os dados

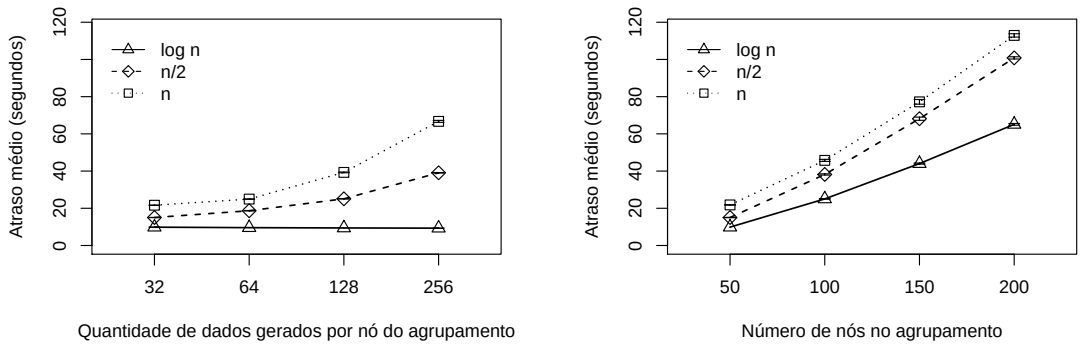
o nível de redução aplicado, menor o consumo de energia.

A segunda análise é realizada variando-se o número de nós por agrupamento na rede e os resultados são apresentados na figura 21(b). Para essa análise, variou-se o número de nós em $s = \{50, 100, 150, 200\}$, fixou-se ainda a quantidade de dados gerados por nó do agrupamento em $n = 32$ e o número de agrupamentos em 2. Nesse caso, quando aumentou-se o número de nós no agrupamento, a energia consumida aumentou consideravelmente. Isso pode ser explicado pelo fato de que o tráfego dentro de cada agrupamento aumenta, uma vez que existem mais nós monitorando o ambiente e, consequentemente, o consumo de energia global da rede também aumenta. Entretanto, quando aplicou-se as reduções $\log_2 n$ e $n/2$, o consumo é menor, especialmente no caso da redução $\log_2 n$, pois a quantidade de pacotes enviados ao sorvedouro é bastante reduzida.

Por fim, avaliou-se nesse cenário, o consumo de energia variando-se o número agrupamentos na rede, e os resultados são mostrados na figura 21(c). Para isso, variou-se o número de

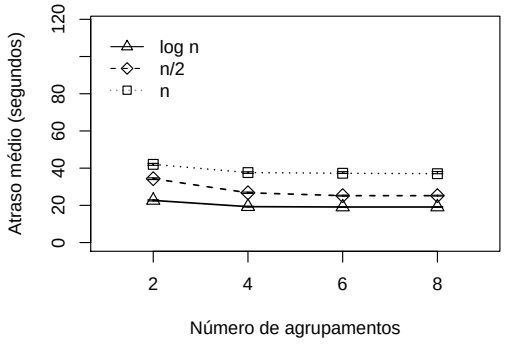
agrupamentos em $\{2, 4, 6, 8\}$, com um número fixo de nós por agrupamento $s = 50$. Fixou-se ainda a quantidade de dados gerados por nó do agrupamento em $n = 32$. Quando variou-se o número de agrupamentos, o consumo de energia variou pouco e, mais uma vez, quanto maior o nível de redução aplicado, ou seja, quanto menor a quantidade de dados enviados pelo nó líder ao sorvedouro, menor o consumo de energia na rede.

A segunda avaliação da redução no nó líder se refere ao atraso médio na entrega dos dados ao sorvedouro. Para essa avaliação, foram considerados os mesmos cenários descritos na avaliação do consumo de energia. Os resultados das simulações são apresentados na figura 22.



(a) Variando o tamanho dos dados enviados pelos nós do agrupamento ao nó líder

(b) Variando número de nós por agrupamento



(c) Variando número de agrupamentos

FIGURA 22: Avaliação do atraso médio na rede ao reduzir os dados

A primeira análise, mostrada na figura 22(a), se refere à simulação feita variando-se o tamanho dos dados enviados pelos nós do agrupamento ao nó líder. Nesse caso, variou-se esse tamanho em $n = \{32, 64, 128, 256\}$, utilizando um número fixo de nós por agrupamento em $s = 50$ e a quantidade de agrupamentos na rede em 2. Nessa avaliação, quando aumentou-

se a quantidade de dados enviados ao nó líder, o atraso também aumentou, exceto quando aplicou-se a redução $\log_2 n$, pois a quantidade de tráfego na rede é muito pequena. Assim como na avaliação do consumo de energia, quando diminui-se a quantidade de dados enviados, o atraso na entrega dos pacotes ao sorvedouro também diminui consideravelmente, uma vez que a quantidade de pacotes transmitidos é menor.

A segunda análise é realizada variando-se o número de nós por agrupamento na rede e os resultados são apresentados na figura 22(b). Para essa análise, variou-se o número de nós em $s = \{50, 100, 150, 200\}$, fixou-se ainda a quantidade de dados gerados por nó do agrupamento em $n = 32$ e o número de agrupamentos em 2. Nesse caso, é possível perceber a mesma relação observada para o consumo de energia, ou seja, à medida que a quantidade de nós por agrupamento é aumentada, o atraso no envio dos pacotes também aumenta consideravelmente, uma vez que o tráfego na rede também é aumentado. Entretanto, quanto menor a quantidade de dados enviados ao sorvedouro, ou seja, quanto maior o nível de redução aplicado, menor o atraso na entrega dos pacotes na rede.

Avaliou-se ainda nesse cenário, o atraso variando-se o número agrupamentos na rede, e os resultados são mostrados na figura 22(c). Para isso, variou-se o número de agrupamentos em $\{2, 4, 6, 8\}$, com um número fixo de nós por agrupamento $s = 50$. Fixou-se ainda a quantidade de dados gerados por nó do agrupamento em $n = 32$. Novamente os resultados mostram a mesma relação observada na avaliação do consumo de energia. Nesse caso, à medida que o número de agrupamentos na rede aumenta, o atraso varia pouco e, mais uma vez, quanto maior o nível de redução aplicado, ou seja, quanto menor a quantidade de dados enviados pelo nó líder ao sorvedouro, menor o atraso na entrega dos pacotes na rede.

6.3 Conclusões parciais

Considerando os resultados apresentados, assim como na avaliação da representatividade dos dados, a técnica proposta se mostrou eficiente no que se refere ao comportamento da rede. Resultados mostraram que ao se utilizar o algoritmo MuSA, o consumo de energia foi consideravelmente reduzido, tanto na redução no sensoriamento, quanto no nó líder, o que resulta em um prolongamento do tempo de vida da rede. Além disso, o atraso no envio dos pacotes na rede também foi reduzido significativamente. Esses resultados reforçam a eficiência do método proposto e comprovam a viabilidade de sua utilização para efetuar a redução de dados multivariados em redes de sensores sem fio.

7 CONCLUSÃO E TRABALHOS FUTUROS

Redes de sensores sem fio possuem restrições de energia, e a extensão do seu tempo de vida é um dos mais importantes problemas no projeto de tais redes. Neste trabalho, foi apresentado o algoritmo de amostragem para redução de dados multivariados, o MuSA. O algoritmo utiliza técnicas baseadas em análise de componentes para classificar os dados, criando uma classificação ordenada, que permite ao MuSA selecionar uma amostra contendo apenas os dados mais relevantes para a aplicação. Dessa forma, a redução da quantidade de dados que trafegam nas redes de sensores sem fio é executada mantendo a representatividade dos dados, diminuindo o atraso e o consumo de energia na rede.

Os resultados mostraram a eficiência do método proposto no que se refere à representatividade dos dados reduzidos. A técnica MuSA se mostrou eficiente em todos os cenários avaliados, obtendo valores satisfatórios tanto para a análise da variância dos dados, como do erro absoluto relativo, inclusive quando a variação dos dados foi considerável, como na avaliação dos dados com ruído. Nesse caso, pode-se dizer que a partir do conjunto de regras $\mathcal{R}'$, é possível tomar um conjunto de decisões D', que são correspondentes ao conjunto de decisões D que seriam tomadas caso todos os dados sensoriados fossem transmitidos. Comparando as técnicas de análise de componentes avaliadas, as amostragens baseadas em PCA e PCA-robusta se mostraram, na maioria dos casos, mais satisfatórias que a baseada em ICA, principalmente na redução no sensoriamento. É importante destacar que é possível que a amostragem baseada em ICA obtenha resultados mais significativos caso seja feita a ordenação das componentes independentes e também caso os dados coletados pelos sensores sejam não gaussianos, cenários não abordados neste trabalho.

A aplicabilidade do MuSA foi observada tanto para a redução no sensoriamento quanto para a redução no nó líder. Entretanto, para a redução no nó líder, os resultados foram ainda mais significativos, uma vez que os erros encontrados foram pequenos inclusive na avaliação com ruído nos dados originais. Por exemplo, considerando a avaliação com dados pseudo-reais com ruído, situação onde a variação dos dados foi mais significativa, os maiores erros observados foram de aproximadamente 15%, aplicando-se a redução $\log_2 n$, o que pode ser considerado satisfatório, especialmente pela pequena quantidade de dados transmitidos ao sorvedouro. Em todos os cenários avaliados, não foram observadas diferenças estatisticamente significativas entre o conjunto de dados original e o conjunto reduzido. Com a redução $n/2$, os

erros foram ainda menores, sendo próximos de zero muitas vezes, o que ratifica a aplicabilidade da técnica proposta. Além disso, a técnica proposta se mostrou escalável em termos da quantidade de dados sensoriados, uma vez que mesmo aumentando a quantidade de dados gerados, os erros observados mantiveram-se aceitáveis e em muitos casos, diminuíram.

A técnica proposta obteve bons resultados também na avaliação do comportamento da rede. Tanto na redução no sensoriamento, quanto no nó líder, a redução dos dados utilizando MuSA resultou em considerável economia de energia na rede. Além disso, o atraso no envio dos pacotes na rede também foi diminuído significativamente em ambos os casos, o que comprova a viabilidade de utilização da técnica proposta neste trabalho para efetuar a redução de dados multivariados em redes de sensores sem fio, mesmo em aplicações que necessitam de uma precisão dos dados alta.

No decorrer do desenvolvimento deste trabalho tivemos como resultado a publicação de um artigo. O artigo *Multivariate Reduction in Wireless Sensor Networks* foi publicado no *14th IEEE Symposium on Computers and Communications* (ISCC'09).

Como trabalhos futuros, pretende-se aplicar o método proposto para processar os dados juntamente com a tarefa de roteamento. Dessa forma, não somente os dados de uma fonte podem ser reduzidos, mas dados similares de diferentes fontes podem ser sujeitos a reduções similares, resultando em maior eficiência de energia. Pretende-se ainda avaliar a redução após realizar a ordenação das componentes independentes e considerando fenômenos descritos por distribuições não gaussianas, com o objetivo de verificar se a técnica ICA possui desempenho melhor nessas condições. Outro aspecto a ser tratado é a análise da solução em outros cenários, onde as perdas dos dados podem afetar a qualidade dos mesmos e também a comparação da técnica proposta com outras técnicas para processamento de dados multivariados, que não são baseadas em análise de componentes.

REFERÊNCIAS

AKYILDIZ, I. F. et al. A survey on sensor networks. *IEEE Communications Magazine*, v. 40, n. 8, p. 102–114, August 2002.

ALBUQUERQUE, E. L. *Compostos Orgânicos Voláteis na Atmosfera Urbana da Região Metropolitana de São Paulo*. Tese (Doutorado) — Universidade Estadual de Campinas, Faculdade de Engenharia Química, May 2007.

ANISI, M. H.; REZAZADEH, J.; DEHGHAN, M. Feda: fault-tolerant energy-efficient data aggregation in wireless sensor networks. In: *16th International Conference on Software, Telecommunications and Computer Networks (SoftCOM'08)*. Split - Dubrovnik, Croatia: IEEE Communication Society, 2008. p. 188–192.

AQUINO, A. L. L. *Redução de Dados em Redes de Sensores Sem Fio Baseada em Stream de Dados*. Tese — Universidade Federal de Minas Gerais, Pós Graduação em Ciência da Computação, February 2008.

AQUINO, A. L. L. et al. Data stream based algorithms for wireless sensor network applications. In: *21st IEEE International Conference on Advanced Information Networking and Applications (AINA'07)*. Niagara Falls, Canada: IEEE Computer Society, 2007. p. 869–876.

AQUINO, A. L. L. et al. Sensor stream reduction for clustered wireless sensor networks. In: *23rd ACM Symposium on Applied Computing 2008 (SAC'08)*. Fortaleza, Brazil: ACM, 2008. p. 2052–2056.

ARAMPATZIS, T.; LYGEROS, J.; MANESIS, S. A survey of applications of wireless sensors and wireless sensor networks. In: *13th IEEE Mediterranean Conference on Control and Automation (MED'05)*. Hawaii, USA: IEEE Computer Society, 2005. p. 719–724.

AZZALINI, A.; CAPITANIO, A. Statistical applications of the multivariate skew normal distribution. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, 1999 Royal Statistical Society, University of Padua, Italy, ; University of Bologna, Italy, v. 61, n. 3, p. 579–602, 1999. ISSN 0047-259X.

AZZALINI, A.; VALLE, A. D. The multivariate skew-normal distribution. *Biometrika*, v. 83, n. 4, p. 715–726, December 1996. Disponível em: <http://biomet.oxfordjournals.org/cgi/content/abstract/83/4/715>.

BUSTOS, O. H.; FRERY, A. C. Reporting monte carlo results in statistics: suggestions and an example. *Revista de la Sociedad Chilena de Estadística*, v. 9, n. 2, p. 46–95, December 1992.

CHATZIGIANNAKIS, V.; PAPAVASSILIOU, S. Diagnosing anomalies and identifying faulty nodes in sensor networks. *IEEE Sensors Journal*, v. 7, n. 5, p. 637–645, May 2007. ISSN 1530-437X.

COMON, P. Independent component analysis, a new concept? *Signal Processing*, Elsevier North-Holland, Inc., Amsterdam, The Netherlands, The Netherlands, v. 36, n. 3, p. 287–314, April 1994. ISSN 0165-1684.

CVEJIC, N.; BULL, D.; CANAGARAJAH, N. Improving fusion of surveillance images in sensor networks using independent component analysis. *IEEE Transactions on Consumer Electronics*, v. 53, n. 3, p. 1029–1035, August 2007. ISSN 0098-3063.

DASGUPTA, K.; KALPAKIS, K.; NAMJOSHI, P. Improving the lifetime of sensor networks via intelligent selection of data aggregation trees. In: *Communication Networks and Distributed Systems Modeling and Simulation Conference (CNDS'03)*. Orlando, Florida, USA: Springer Berlin/Heidelberg, 2003. (Lecture Notes in Computer Science, v. 3397/2005), p. 508–517.

ESTRIN, D. et al. Next century challenges: scalable coordination in sensor networks. In: *5th Annual ACM/IEEE International Conference on Mobile Computing and Networking (MobiCom'99)*. New York, USA: ACM, 1999. p. 263–270. ISBN 1-58113-142-9.

FRERY, A. C. et al. Error estimation in wireless sensor networks. In: *23rd ACM Symposium on Applied Computing 2008 (SAC'08)*. Fortaleza, Brazil: ACM, 2008. p. 1923–1928.

GANESAN, D. et al. Coping with irregular spatio-temporal sampling in sensor networks. *ACM SIGCOMM Computer Communication Review*, v. 34, n. 1, p. 125–130, January 2004.

GAREY, M. R.; JOHNSON, D. S. *Computers and intractibility: a guide to the theory of NP-completeness*. 1st. ed. New York: W.H. Freeman and Company, 1979. ISBN 0716710455.

HOTELLING, H. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, v. 24, n. 1, p. 417–441, 498–520, 1933.

HYVÄRINEN, A. Survey on independent component analysis. *Neural Computing Surveys*, v. 2, n. 1, p. 94–128, April 1999. Disponível em: <http://citeseer.ist.psu.edu/hyv99survey.html>.

HYVäRINEN, A.; OJA, E. A fast fixed-point algorithm for independent component analysis. *Neural Computation*, MIT Press, Cambridge, MA, USA, v. 9, n. 7, p. 1483–1492, October 1997. ISSN 0899-7667.

HYVäRINEN, A. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, v. 10, n. 3, p. 626–634, May 1999.

HYVäRINEN, A.; KARHUNEN, J.; OJA, E. *Independent component analysis*. 1. ed. New York: John Wiley & Sons, 2001. ISBN 047140540X.

JACKSON, J. E. *A user's guide to principal components*. 1. ed. [S.l.]: Wiley-Interscience, 2003. ISBN 978-0-471-47134-9.

JAIN, R. K. *The art of computer systems performance analysis: techniques for experimental design, measurement, simulation, and modeling*. [S.l.]: John Wiley & Sons, 1991. ISBN 0471503361.

JUNIOR, O. S. et al. Multivariate reduction in wireless sensor networks. In: *14th IEEE Symposium on Computers and Communications (ISCC'09)*. Sousse, Tunísia: IEEE Computer Society, 2009.

KALPAKIS, K.; DASGUPTA, K.; NAMJOSHI, P. Maximum lifetime data gathering and aggregation in wireless sensor networks. In: *2nd IEEE International Conference on Networking (ICN'02)*. Atlanta, Georgia, USA: IEEE Computer Society, 2002. p. 685–696.

KIMURA, N.; LATIFI, S. A survey on data compression in wireless sensor networks. *International Conference on Information Technology: Coding and Computing*, IEEE Computer Society, Los Alamitos, USA, v. 2, p. 8–13, 2005.

KRISHNAMACHARI, B.; ESTRIN, D.; WICKER, S. The impact of data aggregation in wireless sensor networks. In: *22nd IEEE International Conference on Distributed Computing Systems (ICDCS'02)*. Vienna, Austria: IEEE Computer Society, 2002. p. 575–578.

KRZANOWSKI, W. J. *Recent advances in descriptive multivariate analysis*. 1. ed. Oxford, USA: Oxford University Press, 1995. ISBN 0198522851.

LI, J.; ZHANG, Y. Interactive sensor network data retrieval and management using principal components analysis transform. *Smart Materials and Structures*, v. 15, n. 11, p. 1747–1757, December 2006.

LOUREIRO, A. A. F. et al. Redes de sensores sem fio (Tutorial). In: *XXI Simpósio Brasileiro de Redes de Computadores (SBRC'03)*. Natal, Brazil: [s.n.], 2003. p. 179–226.

MARBINI, A. D.; SACKS, L. E. Adaptive sampling mechanisms in sensor networks. In: *London Communications Symposium (LCS'03)*. London, UK: [s.n.], 2003.

MCCULLOCH, J. et al. Wireless sensor network deployment for water use efficiency in irrigation. In: *Proceedings of the Workshop on Real-world Wireless Sensor Networks (REALWSN'08)*. New York, USA: ACM, 2008. p. 46–50. ISBN 978-1-60558-123-1.

MINGOTI, S. A. Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada. In: _____. [S.l.]: Editora UFMG, 2005. cap. Análise de Componentes Principais, p. 59–97. ISBN 85-7041-451-X.

MINI, R. A. F.; LOUREIRO, A. A. F. Middleware for network eccentric and mobile applications. In: _____. 1. ed. Berlin: Springer, 2009. cap. Energy in Wireless Sensor Networks, p. 3–24. ISBN 3540897062. In: L. E. Rodrigues, H. Miranda, and B. Garbinato.

NAKAMURA, E. F.; LOUREIRO, A. A. F.; FRERY, A. C. Information fusion for wireless sensor networks: Methods, models, and classifications. *ACM Computing Surveys*, ACM, New York, USA, v. 39, n. 3, p. 1–55, 2007. ISSN 0360-0300.

NAKAMURA, E. F. et al. Using information fusion to assist data dissemination in wireless sensor networks. *Telecommunication Systems*, v. 30, n. 1/2/3, p. 237–254, December 2005. ISSN 1018-4864.

NAKAMURA, E. F. et al. A reactive role assignment for data routing in event-based wireless sensor networks. *Computer Networks*, 2009.

PATTEM, S.; KRISHNAMACHARI, B.; GOVINDAN, R. The impact of spatial correlation on routing with compression in wireless sensor networks. *ACM Transactions on Sensor Network*, ACM, New York, USA, v. 4, n. 4, p. 1–33, 2008. ISSN 1550-4859.

PEARSON, K. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, v. 2, n. 6, p. 559–572, 1901.

POMPILI, D.; MELODIA, T.; AKYILDIZ, I. F. Deployment analysis in underwater acoustic wireless sensor networks. In: *1st ACM International Workshop on Underwater Networks (WUWNet'06)*. New York, USA: ACM, 2006. p. 48–55. ISBN 1-59593-484-7.

ROY, O.; VETTERLI, M. Dimensionality reduction for distributed estimation in the infinite dimensional regime. *IEEE Transactions on Information Theory*, v. 54, n. 4, p. 1655–1669, April 2008. ISSN 0018-9448.

ROYER, E. M.; TOH, C. K. A review of current routing protocols for ad-hoc mobile wireless networks. *IEEE Personal Communications*, v. 6, n. 2, p. 46–55, April 1999.

SANTINI, S.; ROMER, K. An adaptive strategy for quality-based data reduction in wireless sensor networks. In: *3rd International Conference on Networked Sensing Systems (INSS'06)*. Chicago, IL, USA: [s.n.], 2006. p. 29–36.

SEO, S.; KANG, J.; RYU, K. H. Multivariate stream data reduction in sensor network applications. In: *2nd International Symposium on Ubiquitous Intelligence and Smart Worlds (UISW'05)*. Nagasaki, Japan: Springer, 2005. p. 198–207.

SOHRABI, K. et al. Protocols for self–organization of a wireless sensor network. *IEEE Personal Communications*, v. 7, n. 5, p. 16–27, October 2000.

SONG, W.; SHAOWEI, X. Robust PCA based on neural networks. In: . [S.l.: s.n.], 1997. v. 1, p. 503–508.

THOMSON, N. Understanding ANOVA the APL way. *ACM SIGAPL – APL Quote Quad*, ACM, New York, USA, v. 24, n. 1, p. 295–303, August 1993. ISSN 0163-6006.

TILAK, S.; ABU-GHAZALEH, N. B.; HEINZELMAN, W. A taxonomy of wireless micro-sensor network models. *ACM SIGMOBILE Mobile Computing and Communications Review*, v. 6, n. 2, p. 28–36, April 2002.

TORRE, F. De la; BLACK, M. J. Robust principal component analysis for computer vision. In: . [S.l.: s.n.], 2001. v. 1, p. 362–369.

VILLAS, L. A. et al. Um algoritmo de roteamento ciente de agregação de dados para redes de sensores sem fio. In: *XXVII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC'09)*. Recife, Brazil: [s.n.], 2009.

WANG, X.; WANG, S. Collaborative signal processing for target tracking in distributed wireless sensor networks. *Journal of Parallel and Distributed Computing*, Academic Press, Orlando, USA, v. 67, n. 5, p. 501–515, 2007. ISSN 0743-7315.

WILLETT, R.; MARTIN, A.; NOWAK, R. Backcasting: adaptive sampling for sensor networks. In: *3rd International Symposium on Information Processing in Sensor Networks (IPSN'04)*. New York, USA: ACM, 2004. p. 124–133. ISBN 1-58113-846-6.

XU, L.; YUILLE, A. Robust principal component analysis by self-organizing rules based on statistical physics approach. *IEEE Transactions on Neural Networks*, v. 6, n. 1, p. 131–143, January 1995. ISSN 1045-9227.

YU, Y.; PRASANNA, V. K. Energy-balanced task allocation for collaborative processing in wireless sensor networks. *Mobile Networks and Applications*, Springer, v. 10, n. 1-2, p. 115–131, 2005. ISSN 1572-8153.

ZARZOSO, V.; COMON, P.; KALLEL, M. How fast is fastICA? In: *14th European Signal Processing Conference (EUSIPCO'06)*. Florence, Italy: [s.n.], 2006.

ZHANG, Y. et al. Maximum-energy shortest path tree for data aggregation in wireless sensor networks. In: *3rd International Conference on Wireless Communications, Networking and Mobile Computing (WiCom'07)*. Shangai, China: IEEE Computer Society, 2007. p. 2779–2782.

ZHU, J.; PAPAVASSILIOU, S. A resource adaptive information gathering approach in sensor networks. In: *IEEE Sarnoff Symposium on Advances in Wired and Wireless Communication (SARNOFF'04)*. Princeton, USA: IEEE Computer Society, 2004. p. 115–118.

8 ANEXO

Neste capítulo, são apresentados os resultados da avaliação da representatividade dos dados gerados com a distribuição *skew-normal*. Conforme mencionado no capítulo 5, esses resultados são apresentados aqui pelo fato de que, na maioria dos casos, não houve diferenças significativas entre as avaliações com as distribuições *normal* e *skew-normal*. Considerou-se para essa avaliação os mesmos cenários descritos no capítulo 5, ou seja, foram utilizados conjuntos de dados sintéticos e outros pseudo-reais, gerados com e sem a presença de ruído, efetuando a redução Ψ no sensoriamento e no nó líder.

8.1 Dados sintéticos

8.1.1 Redução no sensoriamento

8.1.1.1 Geração dos dados sem ruído

O primeiro cenário a ser avaliado considera a redução no momento do sensoriamento, cuja geração dos dados é realizada sem presença de ruído. Esse cenário segue as mesmas considerações da geração utilizando a distribuição *normal*. Assim, fixou-se o número de sensores em um nó da rede em $s = 5$, com $\mu = \{10, 30, 50, 70, 90\}$, e $\sigma = 10\%$. O tamanho dos dados gerados foi variado em $n = \{256, 512,\ 1024, 2048\}$ e aplicou-se as reduções $n/2$ e $\log_2 n$.

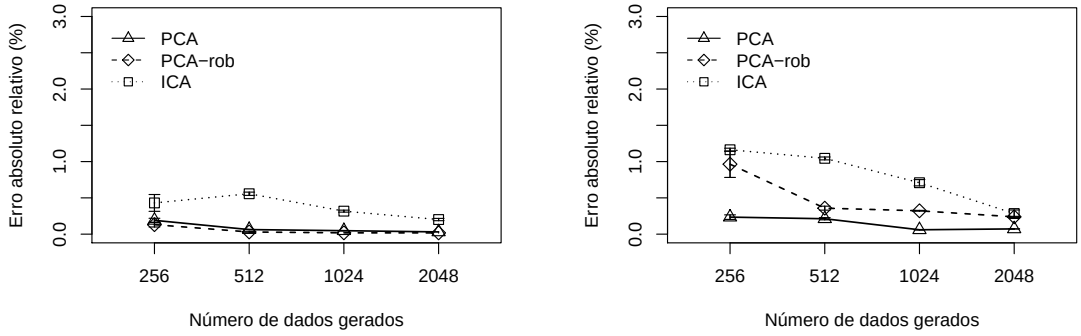
A primeira análise de representatividade dos dados considera $\mathcal{R}'_{\Phi}$ e os valores observados para ambas as técnicas são apresentados na tabela 11. Como na avaliação com a distribuição *normal*, esses resultados indicam que não existem diferenças significativas entre as variâncias dos conjuntos de dados original $\mathcal{V}$ e reduzidos $\mathcal{V}'$, tanto com a utilização da amostragem baseada em PCA quanto em ICA e PCA-robusta. Com isso, para a utilização de dados sem ruído, as decisões D' referentes a $\mathcal{R}'_{\Phi}$ podem ser tomadas.

A segunda análise considera o $\mathcal{R}'_{\Upsilon}$, cujos resultados podem ser observados na figura 23. Assim como na avaliação com a distribuição *normal*, os resultados foram satisfatórios tanto para a redução $n/2$ quanto para a redução $\log_2 n$, e as técnicas apresentaram resultados similares, com uma pequena superioridade de PCA e PCA-robusta. Com isso, para a utilização

TABELA 11: Análise da variância utilizando dados sintéticos sem ruído com redução no sensoriamento

Distribuição	$(n = 256)$		$(n = 512)$		$(n = 1024)$		$(n = 2048)$	
skew-normal	$n/2$	$\log_2 n$	$n/2$	$\log_2 n$	$n/2$	$\log_2 n$	$n/2$	$\log_2 n$
PCA	0,98	0,85	0,98	0,86	0,98	0,88	0,98	0,88
ICA	0,89	0,84	0,84	0,83	0,79	0,84	0,71	0,83
PCA-rob	0,98	0,85	0,98	0,86	0,98	0,87	0,98	0,87

de dados sem ruído, as decisões D' referentes a $\mathcal{R}'_{\Upsilon}$ também podem ser tomadas.



(a) Redução $n/2$ - $\mathcal{V}'^{s}_{n/2}$

(b) Redução $\log_2 n$ - $\mathcal{V}'^{s}_{\log_2 n}$

FIGURA 23: $\mathcal{R}'_{\Upsilon}$ para redução no sensoriamento usando dados sintéticos sem ruído

8.1.1.2 Geração dos dados com ruído

A avaliação a seguir considera a redução no sensoriamento, na qual a geração dos dados é feita com a presença de ruído. Para isso, um ruído aleatório é inserido através do programa R durante a geração dos dados nas distribuição *normal*. Nesse caso, a geração dos dados foi feita a partir das mesmas cinco médias $\mu = \{10, 30, 50, 70, 90\}$, com $\sigma = 10\%$. O tamanho dos dados foi variado em $n = \{256, 512, 1024, 2048\}$ e aplicou-se as reduções $n/2$ e $\log_2 n$.

A primeira avaliação nesse cenário considera o $\mathcal{R}'_{\Phi}$. Valores observados para ambas as técnicas avaliadas são mostradas na tabela 12. Nessa avaliação, as amostragens baseadas em PCA, ICA e PCA-robusta tiveram resultados muito similares em todas as situações simuladas e esses resultados indicam que não existem diferenças significativas entre as variâncias dos conjuntos de dados original e reduzido para as duas técnicas. Com isso, para a utilização de dados com ruído, as decisões D' referentes a $\mathcal{R}'_{\Phi}$ podem ser tomadas.

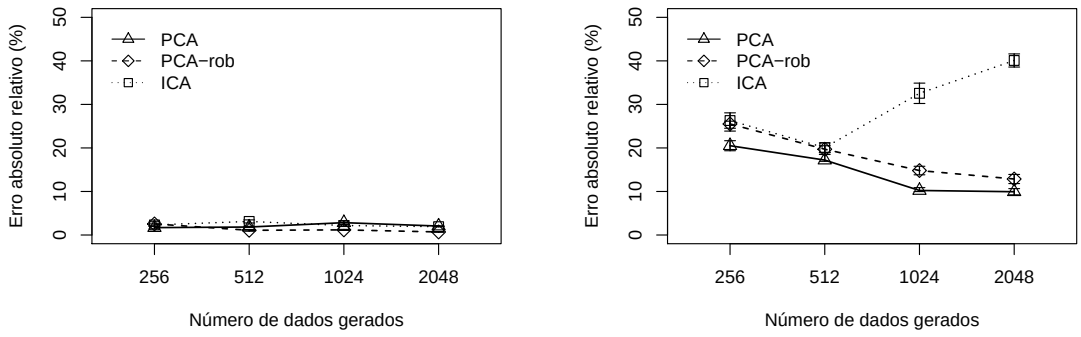
A segunda análise considera o $\mathcal{R}'_{\Upsilon}$ e os resultados podem ser observados na figura 24. Considerando a redução $n/2$ (Figure 24(a)), mesmo com a presença de ruído nos dados, os

TABELA 12: Análise da variância utilizando dados sintéticos com ruído para redução no sensoriamento

Distribuição	$(n = 256)$		$(n = 512)$		$(n = 1024)$		$(n = 2048)$	
skew-normal	$n/2$	$\log_2 n$	$n/2$	$\log_2 n$	$n/2$	$\log_2 n$	$n/2$	$\log_2 n$
PCA	0,39	0,61	0,28	0,54	0,20	0,60	0,08	0,51
ICA	0,39	0,63	0,28	0,55	0,19	0,59	0,08	0,51
PCA-rob	0,35	0,56	0,25	0,58	0,18	0,62	0,08	0,56

resultados observados foram satisfatórios, com erros menores que os observados na distribuição *normal*. Novamente, os resultados mais significativos foram observados quando empregou-se a técnica PCA-robusta e os resultados onde o MuSA se mostrou menos satisfatório foram observados com a técnica ICA.

Considerando a redução $\log_2 n$ (figura 24(b)), os erros observados foram bastante superiores aos da redução $n/2$ e também superiores aos obtidos com a distribuição *normal*. Nesse caso, os melhores resultados foram observados com a amostragem baseada em PCA e novamente a técnica ICA resultado menos satisfatório. Quando empregou-se as técnicas PCA e PCA-robusta, o algoritmo MuSA se mostrou mais satisfatório e à medida que se aumentou a quantidade dos dados gerados $\mathcal{V}$, o $\mathcal{R}'_\Upsilon$ diminuiu em todos os casos, o que comprova a escalabilidade do algoritmo proposto em relação ao número de dados sensoriados, quando a amostragem é baseada nessas duas técnicas. Com isso, para a utilização de dados com ruído, as decisões D' referentes a $\mathcal{R}'_\Upsilon$ podem ser tomadas, especialmente quando da aplicação da redução $n/2$.



(a) Redução $n/2$ - $\mathcal{V}'^{s}_{n/2}$

(b) Redução $\log_2 n$ - $\mathcal{V}'^{s}_{\log_2 n}$

FIGURA 24: $\mathcal{R}'_\Upsilon$ para redução no sensoriamento usando dados sintéticos com ruído

8.1.2 Redução no nó líder

8.1.2.1 Geração dos dados sem ruído

O segundo cenário considera a redução no nó líder. Assim como na análise com a distribuição *normal*, considera-se o monitoramento de um único fenômeno, com $\mu = 50$, e $\sigma = 10\%$. Para as simulações, fixou-se o número de sensores do agrupamento em $s = 400$ e variou-se o tamanho dos dados gerados em $n = \{256, 512, 1024, 2048\}$, aplicando-se as reduções $n/2$ e $\log_2 n$.

Assim como no caso da redução no sensoriamento, a primeira avaliação considera o $\mathcal{R}'_\Phi$ e os resultados são apresentados na tabela 13. Os resultados obtidos mostram que as amostragens baseadas em PCA, ICA e PCA-robusta não apresentam diferenças significativas entre as variâncias dos dados originais e dos dados reduzidos. Com isso, para a utilização de dados sem ruído, as decisões D' referentes a $\mathcal{R}'_\Phi$ podem ser tomadas.

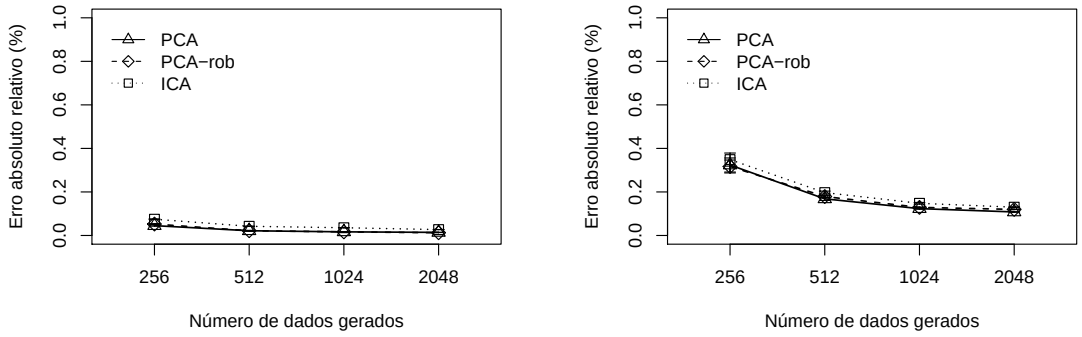
TABELA 13: Análise da variância utilizando dados sintéticos sem ruído com redução no nó líder

Distribuição	$(n = 256)$		$(n = 512)$		$(n = 1024)$		$(n = 2048)$	
skew-normal	$n/2$	$\log_2 n$	$n/2$	$\log_2 n$	$n/2$	$\log_2 n$	$n/2$	$\log_2 n$
PCA	0,90	0,83	0,88	0,81	0,91	0,83	0,89	0,83
ICA	0,83	0,82	0,82	0,82	0,78	0,81	0,77	0,79
PCA-rob	0,89	0,83	0,88	0,84	0,89	0,84	0,89	0,81

A segunda avaliação considera o $\mathcal{R}'_\Upsilon$. Na figura 25, apresenta-se os resultados obtidos para o algoritmo MuSA, com as técnicas PCA, ICA e PCA-robusta. Considerando a a redução $n/2$, mostrada na figura 25(a), o algoritmo se mostrou muito satisfatório com as três técnicas, sendo o $\mathcal{R}'_\Upsilon$ muito próximo a zero em todos os casos, como na avaliação com a distribuição *normal*. Nesse cenário, os resultados observados para as três técnicas foram novamente similares, e o maior $\mathcal{R}'_\Upsilon$ observado foi de aproximadamente $0,1\%$, com a amostragem baseada em ICA. Apesar da pequena diferença entre as técnicas, a amostragem baseada em PCA obteve melhor resultado na maioria dos casos.

Para as simulações utilizando a redução $\log_2 n$, mostrada na figura 25(b), os resultados observados comprovam a aplicabilidade do algoritmo MuSA com ambas as técnicas avaliadas. Nesse caso, assim como na redução $n/2$, os erros observados foram próximos a zero em todos os casos, sendo o maior valor encontrado para o $\mathcal{R}'_\Upsilon$ de aproximadamente $0,4\%$, quando a amostragem foi baseada em ICA e foram gerados $\mathcal{V}^s_{256}$ dados. Como na avaliação com a distribuição *normal*, à medida que aumentou-se a quantidade de dados gerados, o $\mathcal{R}'_\Upsilon$ diminuiu em todos os casos, o que demonstra a escalabilidade do algoritmo MuSA em relação ao número

de dados sensoriados. Nesse caso, embora a diferença entre as técnicas seja muito pequena, a amostragem baseada em PCA obteve resultados melhores na maioria dos tamanhos de dados gerados. Com isso, para a utilização de dados sem ruído, as decisões D' referentes a $\mathcal{R}'_{\Upsilon}$ podem ser tomadas.



(a) Redução $n/2$ - $\mathcal{V}'^{s}_{n/2}$

(b) Redução $\log_2 n$ - $\mathcal{V}'^{s}_{\log_2 n}$

FIGURA 25: $\mathcal{R}'_{\Upsilon}$ para redução no nó líder usando dados sintéticos sem ruído

8.1.2.2 Geração dos dados com ruído

A avaliação a seguir considera a redução no nó líder, na qual a geração dos dados é feita com a presença de ruído. Para isso, assim como na redução no sensoriamento, um ruído aleatório é inserido através do programa R durante a geração dos dados na distribuição *skew-normal*. Como na avaliação sem ruído, os dados foram gerados com $\mu = 50$ e $\sigma = 10\%$. Para as simulações, fixou-se novamente o número de sensores do agrupamento em $s = 400$ e variou-se o tamanho dos dados gerados em $n = \{256, 512,\ 1024, 2048\}$, aplicando-se as reduções $n/2$ e $\log_2 n$.

A primeira análise nesse cenário considera o $\mathcal{R}'_{\Phi}$ e os valores observado são apresentados na tabela 14. Resultados mostram que as amostragens baseadas nas técnicas PCA, ICA e PCA-robusta não apresentam diferenças significativas entre as variâncias dos dados originais e dos dados reduzidos. Nesse caso, os resultados obtidos foram bastante significativos, mais uma vez pelo fato de que todos os sensores monitoram o mesmo fenômeno. Com isso, para a utilização de dados com ruído, as decisões D' referentes a $\mathcal{R}'_{\Phi}$ podem ser tomadas.

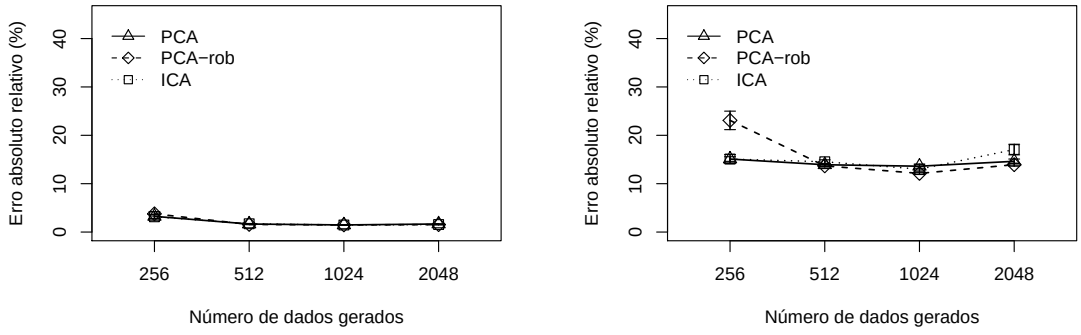
Considera-se agora a avaliação do $\mathcal{R}'_{\Upsilon}$, cujos valores são mostrados na figura 26. Considerando redução $n/2$ (figura 26(a)), os valores observados para as três técnicas foram novamente similares para todos os tamanhos de dados gerados. O algoritmo MuSA se mostrou bastante

TABELA 14: Análise da variância utilizando dados sintéticos com ruído para redução no nó líder

Distribuição	$(n = 256)$		$(n = 512)$		$(n = 1024)$		$(n = 2048)$	
skew-normal	$n/2$	$\log_2 n$	$n/2$	$\log_2 n$	$n/2$	$\log_2 n$	$n/2$	$\log_2 n$
PCA	0,68	0,66	0,61	0,66	0,59	0,60	0,55	0,64
ICA	0,69	0,65	0,60	0,64	0,59	0,63	0,55	0,66
PCA-rob	0,64	0,63	0,64	0,65	0,63	0,64	0,52	0,62

satisfatório para a redução no nó líder, mesmo com a presença de ruído nos dados, o que eleva consideravelmente a variação nos mesmos. Nesse cenário, o maior $\mathcal{R}'_\Upsilon$ observado foi de aproximadamente $3,5\%$, quando utilizou-se a técnica PCA-robusta e gerou-se $\mathcal{V}^s_{256}$ dados.

Para as simulações com a redução $\log_2 n$ (figura 26(b)), os resultados observados com as três técnicas foram novamente similares, exceto quando gerou-se $\mathcal{V}^s_{256}$ dados e utilizou-se a técnica PCA-robusta. Nesse caso, foi encontrado o maior $\mathcal{R}'_\Upsilon$, de aproximadamente 23%. Nos demais casos, os erros observados variaram entre 13% e 17%, aproximadamente, nas três técnicas. Mais uma vez, quando aumentou-se a quantidade de dados gerados, o $\mathcal{R}'_\Upsilon$ diminuiu ou se manteve próximo ao observado com uma menor quantidade de dados, o que ratifica a escalabilidade da solução proposta em termos da quantidade de dados sensoriados, também na redução no nó líder. Com isso, para a utilização de dados com ruído, as decisões D' referentes a $\mathcal{R}'_\Upsilon$ podem ser tomadas.



(a) Redução $n/2$ - $\mathcal{V}'^s_{n/2}$

(b) Redução $\log_2 n$ - $\mathcal{V}'^s_{\log_2 n}$

FIGURA 26: $\mathcal{R}'_\Upsilon$ para redução no nó líder usando dados sintéticos com ruído

8.2 Dados reais

Nesta seção, descreve-se a avaliação do algoritmo MuSA utilizando conjuntos de dados reais de uma aplicação de monitoramento da qualidade do ar. O conjunto de dados reais corresponde às médias de dados coletados durante dois dias, sendo cada valor a média de um fenômeno monitorado em um intervalo de quatro horas. Gerou-se dados pseudo-reais para cada sensor a partir de 12 médias reais μ disponibilizadas por Albuquerque (2007), considerando a distribuição *skew-normal* multivariada. Os dados pseudo-reais representam a estimativa de valores em cada intervalo de quatro horas.

Da mesma forma que a análise com dados sintéticos, considera-se a redução no sensoriamento e no nó líder. Além disso, para cada uma das formas de redução, avaliou-se a representatividade dos dados com e sem a presença de ruído. As avaliações para cada um desses cenários são descritas a seguir.

8.2.1 Redução no sensoriamento

8.2.1.1 Geração dos dados sem ruído

A primeira avaliação considera a redução no momento do sensoriamento, na qual os dados $\mathcal{V}$ são gerados sem a presença de ruído. Nesse caso, simulou-se o monitoramento de 5 diferentes variáveis e, para cada variável, os dados pseudo-reais foram gerados a partir de 12 médias reais μ, com um desvio padrão $\sigma = 10\%$, usando a distribuição *skew-normal*. A quantidade de dados gerados foi variada em $n = \{256, 512, 1024, 2048\}$ e aplicou-se as reduções $n/2$ e $\log_2 n$.

A primeira análise considera os resultados para o $\mathcal{R}'_\Phi$ e os resultados podem ser observados na tabela 15. Nesse caso, as amostragens baseadas em PCA, ICA e PCA-robusta não apresentam diferenças significativas entre as variâncias de seus conjuntos de dados original e reduzido e seus resultados mostram um alto nível de significância. Com isso, para a utilização de dados sem ruído, as decisões D' referentes a $\mathcal{R}'_\Phi$ podem ser tomadas.

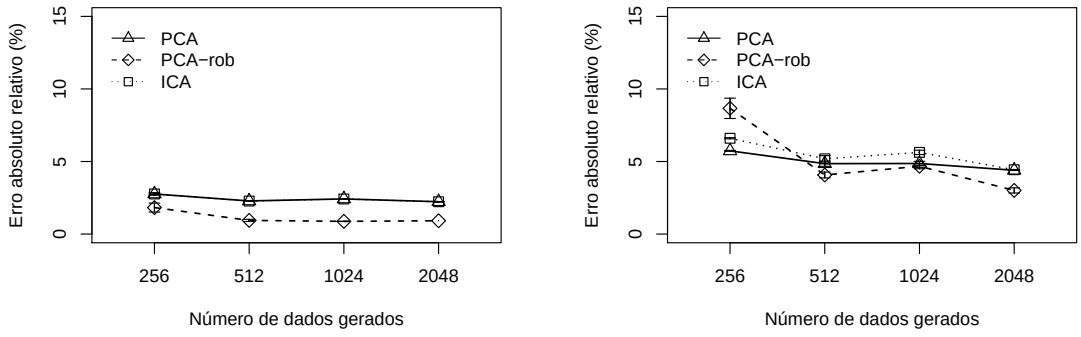
A segunda análise nesse cenário considera o $\mathcal{R}'_\Upsilon$. Resultados de simulações são apresentados na figura 27. No que se refere à redução $n/2$, mostrada na figura 27(a), assim como no avaliação com a distribuição *normal*, o algoritmo MuSA teve resultados satisfatórios com ambas as técnicas utilizadas. Nesse cenário, as técnicas PCA e ICA tiveram resultados quase idênticos e o maior $\mathcal{R}'_\Upsilon$ foi observado para essas técnicas foi de aproximadamente 3%. Os melhores resultados foram novamente obtidos pela técnica PCA-robusta, sendo sua superiori-

TABELA 15: Análise da variância utilizando dados pseudo-reais sem ruído com redução no sensoriamento

Distribuição	$(n = 256)$		$(n = 512)$		$(n = 1024)$		$(n = 2048)$	
skew-normal	$n/2$	$\log_2 n$	$n/2$	$\log_2 n$	$n/2$	$\log_2 n$	$n/2$	$\log_2 n$
PCA	0,82	0,80	0,70	0,77	0,67	0,76	0,61	0,79
ICA	0,81	0,79	0,70	0,75	0,67	0,76	0,61	0,79
PCA-rob	0,75	0,77	0,63	0,74	0,52	0,76	0,39	0,79

dade observada para todos os tamanhos de dados gerados. Mais uma vez ficou comprovada a escalabilidade do MuSA em termos da quantidade de dados sensoriados, uma vez que, com ambas as técnicas utilizadas, quando aumentou-se essa quantidade de dados, o $\mathcal{R}'_\Upsilon$ diminuiu ou se manteve praticamente o mesmo.

Considerando as simulações com a redução $\log_2 n$, como pode ser visto na figura 27(b), a superioridade da amostragem baseada em PCA-robusta ficou comprovada, exceto quando gerou-se $\mathcal{V}^5_{256}$ dados. Assim como na avaliação com a distribuição *normal*, os $\mathcal{R}'_\Upsilon$ observados com a redução $\log_2 n$ foram baixos, sendo próximos aos observados com a redução $n/2$, ratificando a eficiência da técnica MuSA para efetuar a redução considerando o monitoramento de diferentes fenômenos. Nesse caso, o maior $\mathcal{R}'_\Upsilon$ encontrado foi de aproximadamente 8.5%, utilizando a técnica PCA-robusta. Com isso, para a utilização de dados sem ruído, as decisões D' referentes a $\mathcal{R}'_\Upsilon$ podem ser tomadas.



(a) Redução $n/2$ - $\mathcal{V}'^s_{n/2}$

(b) Redução $\log_2 n$ - $\mathcal{V}'^s_{\log_2 n}$

FIGURA 27: $\mathcal{R}'_\Upsilon$ para redução no sensoriamento usando dados pseudo-reais sem ruído

8.2.1.2 Geração dos dados com ruído

A avaliação a seguir retrata a redução no sensoriamento, na qual os dados são gerados com a presença de ruído. Da mesma forma que na avaliação com dados sintéticos, foi introduzido

um ruído aleatório nos dados gerados com a distribuição *skew-normal*, através do programa R. Novamente, simulou-se o monitoramento de 5 diferentes variáveis e, para cada variável, os dados pseudo-reais foram gerados a partir de 12 médias reais μ, com um desvio padrão $\sigma = 10\%$. A quantidade de dados gerados foi variada em $n = \{256, 512, 1024, 2048\}$ e aplicou-se as reduções $n/2$ e $\log_2 n$.

Avaliou-se primeiramente nesse cenário o $\mathcal{R}'_\Phi$ e os resultados obtidos são apresentados na tabela 16. Mais uma vez, os resultados das técnicas avaliadas foram similares com todos os tamanhos de dados gerados e reduções aplicadas. Nesse caso, assim como na avaliação com a distribuição *normal*, o nível de significância dos resultados obtidos é menor, se comparado com os resultados da avaliação sem ruído, o que pode ser explicado pelo significativo aumento da variação dos dados provocada pela introdução do ruído nos mesmos. Entretanto, os resultados ainda podem ser considerados satisfatórios, pois todos os valores estão acima de $0,05$. Com isso, para a utilização de dados com ruído, as decisões D' referentes a $\mathcal{R}'_\Phi$ podem ser tomadas.

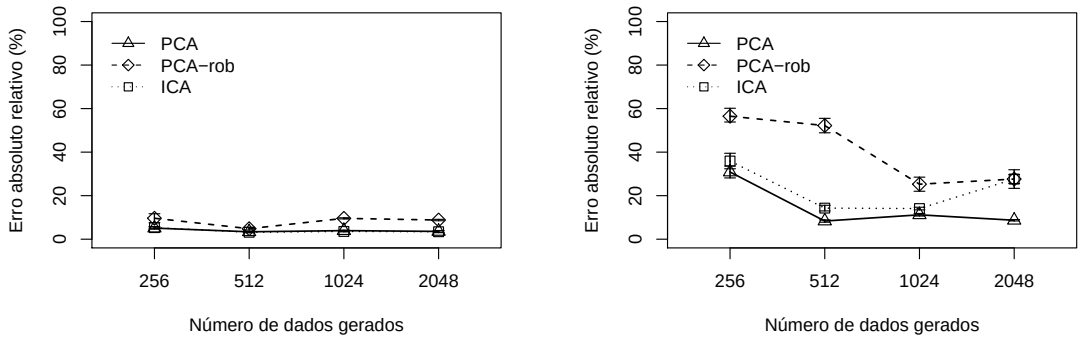
TABELA 16: Análise da variância utilizando dados pseudo-reais com ruído para redução no sensoriamento

Distribuição	$(n = 256)$		$(n = 512)$		$(n = 1024)$		$(n = 2048)$	
skew-normal	$n/2$	$\log_2 n$	$n/2$	$\log_2 n$	$n/2$	$\log_2 n$	$n/2$	$\log_2 n$
PCA	0,39	0,56	0,12	0,53	0,09	0,51	0,07	0,54
ICA	0,39	0,57	0,11	0,52	0,08	0,51	0,07	0,56
PCA-rob	0,36	0,54	0,11	0,52	0,09	0,51	0,07	0,52

A segunda análise nesse cenário considera o $\mathcal{R}'_\Upsilon$, cujos resultados são mostrados na figura 28. Considerando os resultados da avaliação com a redução $n/2$, apresentados na figura 28(a), os melhores resultados foram observados com as amostragens baseadas em PCA e ICA. O $\mathcal{R}'_\Upsilon$ observado com a técnica PCA-robusta foi o maior entre todos casos avaliados, sendo de aproximadamente 10%. É importante destacar que, embora a variação dos dados seja ainda mais considerável devido ao ruído, os resultados obtidos foram satisfatórios, reforçando a viabilidade do uso do MuSA quando da aplicação da redução $n/2$.

No que se refere à redução $\log_2 n$, mostrada na figura 28(b), novamente o algoritmo novamente se mostrou menos satisfatório quando utilizou a técnica PCA-robusta. Com essa técnica, foi observado um $\mathcal{R}'_\Upsilon$ de aproximadamente 58% - o maior entre todos os valores observados. Nesse cenário, o melhor resultado foi observado com a técnica PCA, sendo superior com todos os tamanhos de dados gerados. Nesse cenário, embora os erros observados sejam significativos, os mesmos ainda podem ser considerados satisfatórios, principalmente quando utilizou-se a técnica PCA, uma vez que a variação dos dados foi muito significativa. Com isso, para a utilização de dados com ruído, as decisões D' referentes a $\mathcal{R}'_\Upsilon$ podem ser tomadas,

principalmente quando da utilização da redução $n/2$.



(a) Redução $n/2$ - $\mathcal{V}'^{s}_{n/2}$

(b) Redução $\log_2 n$ - $\mathcal{V}'^{s}_{\log_2 n}$

FIGURA 28: $\mathcal{R}'_{\Upsilon}$ para redução no sensoriamento usando dados pseudo-reais com ruído

8.2.2 Redução no nó líder

8.2.2.1 Geração dos dados sem ruído

A avaliação a seguir diz respeito à redução no nó líder, na qual a geração dos dados se dá sem a presença de ruído. Os dados $\mathcal{V}$ foram gerados representando o sensoriamento de uma única variável por diferentes nós sensores em um agrupamento. Nesse caso, o agrupamento é composto por 400 nós e para cada nó, o dado pseudo-real considera 12 médias reais μ, com $\sigma = 10\%$ e usando a distribuição *skew-normal*. Assim como nas avaliações anteriores, variou-se o tamanho dos dados gerados em $n = \{256, 512, 1024, 2048\}$ e aplicou-se as reduções $n/2$ e $\log_2 n$.

Avaliou-se primeiramente os resultados para o $\mathcal{R}'_{\Phi}$, cujos valores são mostrados na tabela 17. Como na avaliação com dados sintéticos, os resultados mostram que não existem diferenças significativas entre as variâncias de seus conjuntos de dados original $\mathcal{V}$ e reduzido $\mathcal{V}'$ considerando as três técnicas empregadas e, em todos os casos, os resultados apresentaram alto nível de significância. Com isso, para a utilização de dados sem ruído, as decisões D' referentes a $\mathcal{R}'_{\Phi}$ podem ser tomadas.

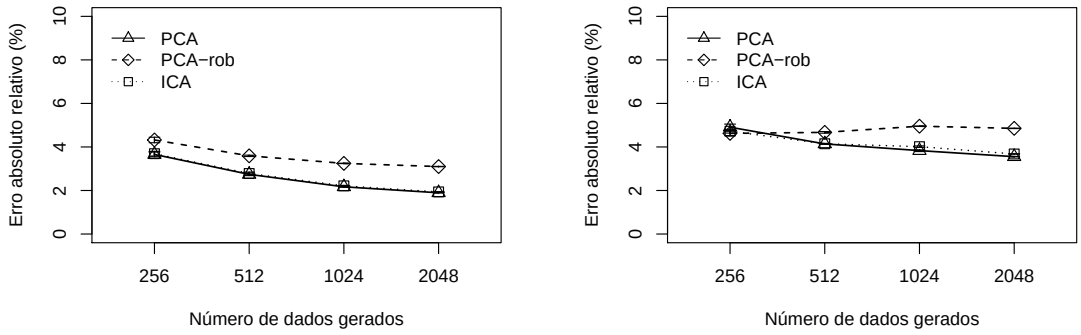
A segunda análise nesse cenário considera o $\mathcal{R}'_{\Upsilon}$ e os resultados observados são apresentados na figura 29. Considerando a redução $n/2$, mostrada na figura 29(a), as técnicas PCA e ICA apresentaram resultados praticamente idênticos para todos os tamanhos de dados gerados, e os valores observados para o $\mathcal{R}'_{\Upsilon}$ mostram a aplicabilidade do algoritmo MuSA nesse

TABELA 17: Análise da variância utilizando dados pseudo-reais sem ruído com redução no nó líder

Distribuição	$(n = 256)$		$(n = 512)$		$(n = 1024)$		$(n = 2048)$	
skew-normal	$n/2$	$\log_2 n$	$n/2$	$\log_2 n$	$n/2$	$\log_2 n$	$n/2$	$\log_2 n$
PCA	0,79	0,76	0,58	0,74	0,65	0,72	0,60	0,77
ICA	0,78	0,75	0,57	0,72	0,63	0,71	0,58	0,76
PCA-rob	0,78	0,76	0,57	0,74	0,64	0,71	0,60	0,76

cenário, especialmente pelo fato de que a variação nos dados originais $\mathcal{V}$ foi considerável. A situação em que o algoritmo se mostrou menos satisfatório foi na utilização de PCA-robusta, sendo o maior $\mathcal{R}'_\Upsilon$ encontrado de aproximadamente $4,5\%$.

Os resultados observados na avaliação da redução $\log_2 n$, mostrados na figura 29(b), foram similares aos observados com a redução $n/2$, o que também reforça a aplicabilidade do MuSA. Como na avaliação com a distribuição *normal*, nesse cenário, o maior $\mathcal{R}'_\Upsilon$ observado foi de aproximadamente 5%, valor muito baixo se considerarmos o tamanho da amostra dos dados $\mathcal{V}'$ que seria enviada ao sorvedouro. Além disso, quando aumentou-se a quantidade de dados gerados, o $\mathcal{R}'_\Upsilon$ diminuiu em todos os casos, mostrando mais uma vez a escalabilidade da solução proposta em termos da quantidade de dados sensoriados. Com isso, para a utilização de dados sem ruído, as decisões D' referentes a $\mathcal{R}'_\Upsilon$ podem ser tomadas.



(a) Redução $n/2$ - $\mathcal{V}'^s_{n/2}$

(b) Redução $\log_2 n$ - $\mathcal{V}'^s_{\log_2 n}$

FIGURA 29: $\mathcal{R}'_\Upsilon$ para redução no nó líder usando dados pseudo-reais sem ruído

8.2.2.2 Geração dos dados com ruído

Discute-se a seguir a avaliação da redução no nó líder, com os dados sendo gerados com através do programa R, com a introdução de um ruído aleatório nesses dados. Assim como na avaliação com os dados sem ruído, gerou-se dados representando o sensoriamento de uma

única variável por diferentes nós sensores em um agrupamento. O agrupamento é composto por 400 nós e para cada nó, o dado pseudo-real considera 12 médias reais μ, com $\sigma = 10\%$ e usando as distribuição *skew-normal*. Mais uma vez, variou-se o tamanho dos dados gerados em $n = \{256, 512, 1024, 2048\}$ e aplicou-se as reduções $n/2$ e $\log_2 n$.

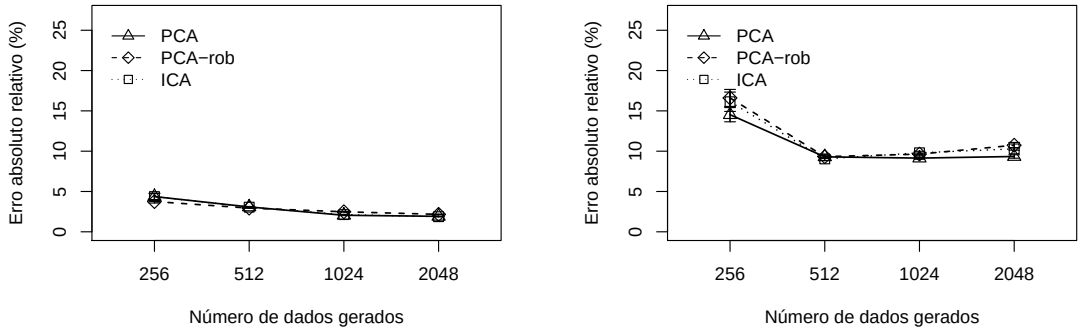
A primeira análise nesse cenário é feita sobre o $\mathcal{R}'_\Phi$, que tem seus resultados apresentados na tabela 18. Como nos cenários anteriores, os resultados mostraram que não existem diferenças significativas entre as variâncias dos conjuntos de dados original $\mathcal{V}$ e reduzido $\mathcal{V}'$ em todos os casos simulados. Com isso, para a utilização de dados com ruído, as decisões D' referentes a $\mathcal{R}'_\Phi$ podem ser tomadas.

TABELA 18: Análise da variância utilizando dados pseudo-reais com ruído para redução no sensoriamento

Distribuição	$(n = 256)$		$(n = 512)$		$(n = 1024)$		$(n = 2048)$	
skew-normal	$n/2$	$\log_2 n$	$n/2$	$\log_2 n$	$n/2$	$\log_2 n$	$n/2$	$\log_2 n$
PCA	0,11	0,55	0,15	0,52	0,49	0,64	0,59	0,64
ICA	0,11	0,59	0,15	0,52	0,49	0,61	0,59	0,64
PCA-rob	0,10	0,57	0,15	0,54	0,43	0,61	0,58	0,66

A segunda análise, cujos resultados são mostrados na figura 30, é feita sobre o $\mathcal{R}'_\Upsilon$. Em relação à redução $n/2$, mostrada na figura 30(a), os resultados observados foram similares aos da avaliação com dados sem ruído. Como na avaliação com a distribuição *normal*, as técnicas PCA, ICA e PCA-robusta apresentaram resultados similares em todos os casos e o maior $\mathcal{R}'_\Upsilon$ encontrado foi de aproximadamente 5%. Ficou também comprovada a escalabilidade da solução em termos da quantidade de dados sensoriados, visto que quando aumentou-se essa quantidade de dados, o $\mathcal{R}'_\Upsilon$ diminuiu em todos os casos.

Considerando a redução $\log_2 n$, que tem seus resultados apresentados na figura 30(b), os resultados obtidos com as três técnicas analisadas foram novamente similares, havendo uma pequena superioridade da amostragem baseada em PCA. Da mesma forma que na análise com a distribuição *normal*, os resultados também mostraram que a técnica MuSA é bastante viável para redução nó líder, mesmo quando a variação nos dados é muito grande, como nesse caso. O maior valor observado para o $\mathcal{R}'_\Upsilon$ foi de aproximadamente 17%, utilizando-se as técnicas ICA e PCA-robusta, com $\mathcal{V}^s_{256}$ dados gerados. Com isso, para a utilização de dados com ruído, as decisões D' referentes a $\mathcal{R}'_\Upsilon$ podem ser tomadas.



(a) Redução $n/2$ - $\mathcal{V}'^{s}_{n/2}$

(b) Redução $\log_2 n$ - $\mathcal{V}'^{s}_{\log_2 n}$

FIGURA 30: $\mathcal{R}'_{\Upsilon}$ para redução no nó líder usando dados pseudo-reais com ruído

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo