Um Algoritmo de Amostragem Multivariada Para Redes de Sensores Sem Fio

Orlando Silva Junior[1], **Andre L. L. Aquino**[2], **Raquel A. F. Mini**[1]

[1] Departamento de Ciência da Computação
Pontifícia Universidade Católica de Minas Gerais (PUC Minas)
Belo Horizonte, MG, Brasil

[2]Departamento de Computação
Universidade Federal de Ouro Preto (UFOP)
Ouro Preto, MG, Brasil

orlando.junior@sga.pucminas.br,alla@iceb.ufop.br,raquelmini@pucminas.br

Abstract. *A wireless sensor network (WSN) is energy constrained, and the extension of its lifetime is an important issue in its design. A WSN can be used to collect and process data of environment that can be, in some cases, multivariate. In this way, to help the data processing, this work proposes a multivariate sampling algorithm which uses component analysis techniques to rank the data and then to select only the data more relevant to the application. Simulation results show that our technique reduces the data keeping its representativeness. In addition, the energy consumption and delay on the network are reduced.*

Resumo. *Uma rede de sensores sem fio (RSSF) possui restrições de energia, e a extensão do seu tempo de vida é uma importante questão no projeto dessas redes. As RSSFs podem ser usadas para coletar e processar dados do ambiente, que podem ser, em alguns casos, multivariados. Dessa forma, para auxiliar o processamento dos dados, este trabalho propõe um algoritmo de amostragem multivariada que usa técnicas de análise de componentes para classificar os dados e então selecionar apenas os mais relevantes para a aplicação. Resultados de simulação mostram que a técnica reduz os dados mantendo sua representatividade. Além disso, o consumo de energia e o atraso na rede são reduzidos.*

1. Introdução

O mundo ao nosso redor possui uma variedade de fenômenos descritos por algumas variáveis, tais como temperatura, pressão e umidade, que podem ser monitoradas por dispositivos com capacidade de sensoriamento, processamento e comunicação. Tais dispositivos trabalhando cooperativamente compõem as redes de sensores sem fio (RSSFs) [Akyildiz et al. 2002]. Uma característica que distingue as RSSFs de outras redes é que os nós sensores têm recursos muito limitados. Os nós que compõem a rede são equipados com bateria e, em diversas aplicações, serão depositados em áreas remotas, impedindo o acesso aos mesmos para manutenção. Nesse cenário, o tempo de vida da rede depende da quantidade de energia disponível no nó sensor, fazendo-se necessário o balanceamento desses recursos limitados para aumentar esse tempo de vida.

Considerando as características dos fenômenos, é importante distinguir os dados sensoriados como univariados ou multivariados. Dados univariados representam uma

amostra de um mesmo tipo de fenômeno, por exemplo, um nó que monitora apenas a temperatura do ambiente. Por outro lado, os dados multivariados representam amostras de diferentes fenômenos. Essas amostras são originadas de diferentes sensores de um nó específico ou do mesmo tipo de sensor de diferentes nós, por exemplo, um nó que possui sensores que monitoram temperatura, umidade e pressão simultaneamente, ou um nó que processa os dados de um conjunto de nós que monitoram somente temperatura.

Baseado nesses aspectos, este trabalho propõe um algoritmo de amostragem multivariada para RSSFs, o MuSA (***Mu**ltivariate **S**ampling **A**lgorithm*). No MuSA, técnicas de análise de componentes são usadas para classificar os dados sensoriados considerando os escores da primeira componente. Com base nessa classificação, a amostragem é efetuada no conjunto de dados original. Dessa forma, duas hipóteses podem ser consideradas: (i) o uso de técnicas de análise de componentes para classificar os dados pode auxiliar na amostragem, mantendo sua representatividade; e (ii) através da amostragem, é possível diminuir o consumo de energia e o atraso no envio das mensagens na rede.

MuSA pode ser usado em diferentes aplicações que geram dados multivariados. Além disso, o algoritmo tem seus parâmetros facilmente ajustados para obter um melhor desempenho na amostragem. Considerando os requisitos da aplicação, é possível escolher como a amostragem será efetuada, por exemplo, usando os escores superiores, inferiores ou intermediários da primeira componente calculada. Se a aplicação está interessada em identificar os valores discrepantes nos dados de entrada, a amostragem pode ser efetuada considerando os escores superiores (interesse nos valores mais altos) ou inferiores (interesse nos valores mais baixos) na primeira componente. Se a aplicação está interessada em identificar os valores considerados "normais", a amostragem deve ser efetuada considerando os escores intermediários.

Este artigo está organizado como segue. A seção 2 mostra os trabalhos relacionados. A seção 3 discute sobre a caracterização da amostragem multivariada em RSSFs. Na seção 4, é apresentado o algoritmo de amostragem multivariada (MuSA) proposto neste trabalho. A seção 5 discorre a respeito dos resultados de simulação referentes à análise da representatividade dos dados. Na seção 6, avalia-se o comportamento da rede usando o algoritmo MuSA. Por fim, na seção 7, apresenta-se a conclusão e os trabalhos futuros.

2. Trabalhos relacionados

Considerando a redução univariada, uma técnica usual é a agregação, na qual cada nó decide se irá ou não agregar os dados, de acordo com seu nível de energia e o tempo para a entrega dos dados [Zhu and Papavassiliou 2004]. Outra abordagem utilizada é a redução baseada em *stream* de dados, na qual o dado sensoriado é caracterizado como um *stream* de dado e uma estratégia de amostragem específica é executada [Aquino et al. 2007]. Outras estratégias empregadas em RSSFs são fusão de dados [Nakamura et al. 2007] e amostragem adaptativa [Santini and Romer 2006].

Considerando dados multivariados, algumas técnicas podem ser descritas. Em [Seo et al. 2005], os autores comparam os métodos DWT (*Discrete Wavelet Transformation*), HCL (*Hierarchical Clustering*), Amostragem and SVD (*Singular Value Decomposition*). Resultados apresentados mostram que a amostragem teve desempenho superior. É importante destacar que, no que diz respeito à amostragem multivariada em RSSFs, não foram encontrados outros trabalhos além desse.

Com relação às técnicas de análise de componentes, exemplos de métodos empregados para processar dados multivariados em RSSF são a Análise de Componentes Principais (PCA - *Principal component analysis*) e Análise de Componentes Independentes (ICA - *Independent component analysis*). Em [Cvejic et al. 2007], é apresentado um algoritmo baseado em ICA para melhorar a fusão de imagens de vigilância. O método combina o uso de PCA e ICA para reduzir os dados. É possível ainda encontrar contribuições que aplicam PCA em conjunto com técnicas de predição para reduzir dados [Li and Zhang 2006]. O principal objetivo do algoritmo é melhorar a transmissão e o gerenciamento de vibração em larga escala em sistemas de monitoramento estrutural. Outro trabalho que usa PCA em RSSFs é apresentado em [Roy and Vetterli 2008]. Nesse caso, PCA é usada para reduzir dados em aplicações de áudio.

Diferentemente dos trabalhos anteriores, o algoritmo proposto utiliza três técnicas de análise de componentes para classificar os dados sensoriados, auxiliando no processo de amostragem. Além disso, outro diferencial é que o algoritmo é flexível em relação à forma de se realizar a classificação, podendo utilizar os escores superiores, inferiores ou intermediários da primeira componente calculada pela técnica de análise de componentes empregada. Assim, de acordo com a aplicação, o algoritmo pode ser facilmente modificado para obter um desempenho mais satisfatório.

3. Amostragem multivariada em RSSFs

O sensoriamento de dados multivariados em RSSFs pode ser modelado de acordo com o diagrama apresentado na figura 1. 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 é completada sem problemas, i.e., sem a possibilidade de perdas de informação, tem-se um conjunto de regras ($\mathcal{R}^*$) ideal para a tomada de uma série de decisões ideais (D^*).

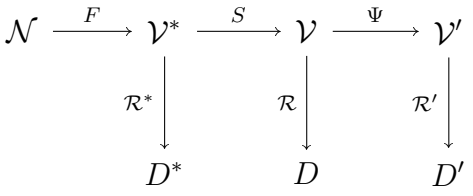


Figura 1. Caracterização da amostragem multivariada em RSSFs

Entretanto, em RSSFs, ao invés de uma situação ideal, tem-se um conjunto de s nós sensores, $S = (S_1, \ldots, S_s)$, no qual cada sensor S_i monitora p fenômenos e produz um conjunto de dados $\mathcal{V}_{1..p,1..n}$, onde n é o número de leituras de cada fenômeno. Usando todos os dados sensoriados, é possível conceber um conjunto de regras ($\mathcal{R}$) para prover um conjunto de decisões (D). Contudo, devido às restrições das RSSFs, enviar grandes quantidades de dados pode ser oneroso em termos de energia, largura de banda e tempo de resposta, causando atraso excessivo e diminuindo o tempo de vida da rede. Dessa forma, usar todo o conjunto $\mathcal{V}$ pode ser inviável e o uso de técnicas de amostragem pode auxiliar na resolução desse problema. Na figura 1, a amostragem multivariada é representada por

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

onde n é a quantidade de dados coletados por cada sensor s e n' o número de dados do conjunto reduzido. Considerando o uso de análise de componentes, Ψ é dada por

$$\Psi = \psi_A \circ \psi_O \circ \psi_C,$$

onde ψ_C é o cálculo das componentes, ψ_O é a classificação dos escores da primeira componente e ψ_A representa a amostragem efetuada. Para efetuar Ψ, diferentes técnicas de análise de componentes (ψ_C) podem ser aplicadas, tais como:

- **Análise de componentes principais (PCA):** realiza uma transformação linear em um conjunto de dados, com o objetivo de obter um novo conjunto de dados com dimensões reduzidas [Pearson 1901].
- **Análise de componentes principais robusta (PCA-robusta):** técnica desenvolvida com o objetivo de tratar possíveis deficiências da técnica PCA tradicional no que se refere à presença de valores discrepantes ou atípicos no dados [Ruymgaart 1981].
- **Análise de componentes independentes (ICA):** também realiza uma transformação linear nos dados, considerando que suas componentes são estatisticamente independentes. Essa técnica pode reduzir, aumentar ou manter a dimensão dos dados originais [Hyvärinen 1999].

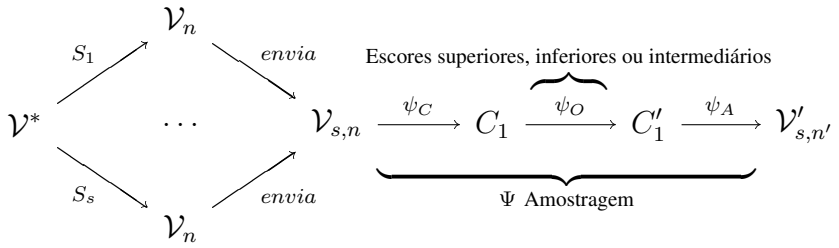
O conjunto de dados reduzido do domínio $\mathcal{V}$ é representado por $\mathcal{V}'_{p,n'}$. Com isso, as novas regras relacionadas a $\mathcal{V}'$ são representadas por $\mathcal{R}'$ e elas levam a um conjunto de decisões D'. Para efetuar Ψ em $\mathcal{V}$, um ponto fundamental a ser analisado é se a partir de $\mathcal{V}'$ e de acordo com $\mathcal{R}'$, é possível tomar as decisões D', sendo D' equivalente a D, que seriam tomadas caso fosse utilizado o conjunto de dados original $\mathcal{V}$.

Em RSSFs, as principais topologias utilizadas são a plana e a hierárquica. Com isso, é preciso definir o comportamento do processo de amostragem Ψ para essas duas topologias. Considerando as redes planas, tem-se o seguinte:

$$\mathcal{V}^* \xrightarrow{S_p} \mathcal{V}_{p,n} \xrightarrow{\psi_C} \overbrace{C_1}^{\text{Escores superiores, inferiores ou intermediários}} \xrightarrow{\psi_O} C'_1 \xrightarrow{\psi_A} \mathcal{V}'_{p,n'}$$
$$\underbrace{\qquad\qquad\qquad\qquad\qquad\qquad}_{\Psi \text{ Amostragem}}$$

Nesse caso, s nós sensores coletam simultaneamente p diferentes fenômenos do ambiente para compor $\mathcal{V}_{p,n}$, onde n representa as diferentes leituras ao longo do tempo. Esses nós executam a amostragem Ψ depois de n leituras, impedindo o tráfego de dados desnecessário na rede. A amostragem Ψ é efetuada da seguinte forma: (i) realiza-se o processamento ψ_C através de PCA, PCA-robusta ou ICA; (ii) é feita a classificação ψ_O dos dados coletados por cada sensor com base na quantidade de dados n' do conjunto reduzido e no tipo de escores da primeira componente a ser empregado, i.e., os escores superiores, inferiores ou intermediários; e (iii) é realizada a amostragem ψ_A, de acordo com o nível de redução desejado, que selecionará os dados identificados como mais relevantes para a aplicação, de forma que as decisões a serem tomadas não sejam comprometidas. Após a amostragem, o conjunto reduzido $\mathcal{V}'_{p,n'}$ é então enviado pelo nó sensor ao nó *sink*.

Considerando uma rede hierárquica, tem-se:



O processo é similar ao ilustrado para redes planas. A principal diferença é que cada nó sensor do agrupamento monitora apenas um fenômeno $\mathcal{V}_n$ e envia os dados coletados para o nó líder, que é responsável por receber esses dados e efetuar a amostragem Ψ sobre $\mathcal{V}_{s,n}$, que é feita da mesma forma apresentada anteriormente para redes planas. Ao final da amostragem dos dados, o nó líder envia o conjunto reduzido $\mathcal{V}'_{s,n'}$ para o *sink*.

4. Algoritmo MuSA

Baseado na caracterização apresentada na seção anterior, é apresentado aqui o algoritmo MuSA – ***Mu**ltivariate **S**ampling **A**lgorithm*. O objetivo do algoritmo é efetuar a redução dos dados com a perda mínima de informações e, consequentemente, diminuir o consumo de energia e o atraso na rede, considerando as topologias plana e hierárquica. Considere os dados de entrada $\mathcal{V}_{\gamma,n}$ ($\gamma = p$ ou $\gamma = s$), onde $n > 0$ representa os valores monitorados por cada sensor e $\gamma \geq 1$ representa os sensores ou os nós sensores que coletam esses dados. Para reduzir $\mathcal{V}$, MuSA executa os seguintes passos:

Passo 1. O conjunto original de dados $\mathcal{V}$ é usado para calcular as componentes C;
Passo 2. A primeira componente C_1 é selecionada e seus escores são ordenados;
Passo 3. As posições dos escores (superiores, inferiores ou intermediários) em C_1 são usadas para identificar as linhas em $\mathcal{V}$ que irão compor o conjunto reduzido $\mathcal{V}'$.
Passo 4. O conjunto reduzido $\mathcal{V}'$, contendo os dados mais relevantes em $\mathcal{V}$ é obtido.

Um pseudo-código simplificado do MuSA é mostrado no algoritmo 1. Considerou-se I os índices de C_1 e J os índices intermediários usados na amostragem.

Algorithm 1 Algoritmo de amostragem multivariada

Require: $\mathcal{V}$ – dado original, n' – tamanho da redução
Ensure: $\mathcal{V}'$ – dado reduzido

1: $C \leftarrow calculaComponentes(\mathcal{V})$ {Pode usar PCA, PCA-robusta, ou ICA}
2: $I \leftarrow ordena(C_1)$
3: $J \leftarrow valoresEscores(I, r)$ {Superiores, inferiores ou intermediários}
4: **for** $i \leftarrow 1$ **to** n' **do**
5: $\quad \mathcal{V}'_{\gamma,i} \leftarrow \mathcal{V}_{\gamma,J_i}$
6: **end for**

Analisando o algoritmo 1:

- Na linha 1, tem-se o cálculo das componentes através da técnica escolhida. A complexidade do cálculo de PCA pode ser estimada em $O(\gamma^2\gamma' + \gamma^2 n)$, onde γ se refere à dimensão dos dados originais (número de sensores ou nós sensores), γ' é a dimensão dos dados reduzidos e n é a quantidade de dados gerados. Se $\gamma > \gamma'$ e

$\gamma > n$, a ordem de complexidade pode ser estimada em $O(\gamma^2)$. Como nesse caso $\gamma = \gamma'$ e $\gamma < n$, tem-se $O(\gamma^2 n)$. Para o cálculo de ICA, considerando o algoritmo FastICA, essa ordem pode ser estimada em $O(\gamma\, n)$. A ordem de complexidade de PCA-robusta pode ser estimada em $O(\gamma\, k^2 n)$, onde k representa o número de componentes principais desejado. Como somente a primeira componente principal é usada, a ordem de complexidade é $O(\gamma\, n)$.

- Na linha 2, a primeira componente C_1 é ordenada, onde os índices (I) dos escores ordenados são obtidos. Sua complexidade, considerando um algoritmo quicksort é $O(n\, \log_2 n)$, uma vez que $|C_1| = n$.
- Na linha 3, são selecionados os (J) valores intermediários em I. Essa seleção pode ser realizada considerando a escala de elementos $[|C_1|/2 - n'/2, |C_1|/2 + n'/2]$.
- Nas linhas 4 – 6, é obtido o conjunto reduzido, com ordem de complexidade $O(n')$.

Sendo assim, a complexidade de tempo total usando PCA é

$$O(\gamma^2 n) + O(n \log_2 n) + O(n' \log_2 n') + O(n') = O(\gamma^2 n).$$

Considerando ICA ou PCA-robusta, essa complexidade é

$$O(\gamma\, n) + O(n \log_2 n) + O(n' \log_2 n') + O(n') = O(\gamma\, n).$$

Para a complexidade de espaço, considere as matrizes $\mathcal{V}$, $\mathcal{V}'$, C e Σ, que representam, respectivamente, os dados de entrada, dados de saída, componentes principais ou independentes e a matriz de covariância dos dados. A complexidade de espaço é dada por

$$3\, O(\gamma\, n) + O(n\, n') = O(\gamma\, n).$$

Uma vez que cada nó fonte envia $\mathcal{V}'$ ao *sink*, a complexidade de comunicação é

$$O(n\, n'\, L),$$

onde L é a maior rota na rede.

5. Avaliação da representatividade dos dados

Todos os cenários de simulação consideraram a Simulação de Monte Carlo, que é um método estatístico usado para realizar simulações estocásticas. Para cada quantidade de sensores avaliada, o algoritmo foi simulado com todos os tamanhos de dados considerados. Todas as simulações foram feitas com o algoritmo implementado no programa estatístico R[1]. O número de simulações necessárias foi calculado de acordo com [Jain 1991] e é dado por

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

onde z é uma constante de valor $1,96$, s_d é o desvio padrão encontrado nas primeiras 5 simulações, $\overline{\mathcal{V}}$ é o valor médio obtido e p_c é a porcentagem da média desejada como desvio, que nesse caso foi de 5%. Com isso, cada cenário foi executado com 93 diferentes conjuntos de dados, com o algoritmo utilizando os escores intermediários.

Considerando o processo de amostragem $\mathcal{V} \rightarrow \mathcal{V}' \rightarrow D'$, apresentado na figura 1, as avaliações foram feitas conforme descrito a seguir:

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

- $\mathcal{V}$ é representado por dados sintéticos e pseudo-reais. Os dados sintéticos são gerados através da distribuição normal multivariada. Além disso, os dados foram gerados com e sem a presença de ruído. No caso dos dados com ruído, eles são primeiramente gerados com a distribuição normal e então é introduzido um ruído aleatório nos mesmos. Nesse caso, os dados não representam um fenômeno específico, sendo utilizados valores arbitrários para as médias.
  Os dados pseudo-reais (extraídos de `http://www.cetesb.sp.gov.br/`) correspondem à média de dados reais coletados durante dois dias, sendo cada valor a média de um fenômeno monitorado em um intervalo de quatro horas. Os dados pseudo-reais foram gerados para cada sensor a partir de 12 médias reais (considerando a distribuição *normal* multivariada). Entre os diferentes fenômenos avaliados, foram considerados como dados pseudo-reais os poluentes n-hexano, metilciclopentano, tolueno, p-xileno e 1,3,5-TMB.
- Com o objetivo de estabelecer possíveis limites para os níveis de redução aplicados, de forma a não comprometer as decisões D' tomadas, foram utilizados dois níveis de redução para compor $\mathcal{V}'$, $n' = n/2$ e $n' = \log_2 n$. A partir da avaliação com esses níveis é possível então estimar os valores adequados em função da aplicação a ser considerada, uma vez que para cada aplicação de interesse, diferentes níveis podem ser aplicados para obter um resultado satisfatório.
- Para avaliar a representatividade dos dados reduzidos em relação aos dados originais, dois testes estatísticos foram utilizados como regras R'. O primeiro foi o teste de hipótese *Analysis of Variance – ANOVA* [Thomson 1993], que tem como objetivo avaliar se existem diferenças estatisticamente significativas entre as médias do conjunto de dados original e do conjunto reduzido. O cálculo é dado por

$$F = \lambda_B^2 / \lambda_W^2,$$

  onde λ_B^2 representa a dispersão entre os conjuntos $\mathcal{V}$ e $\mathcal{V}'$, e λ_W^2 a dispersão dentro dos conjuntos. A partir desse cálculo, o $p - valor$ é usado para determinar se a hipótese nula H_0 deve ser aceita ou rejeitada. Nesse caso, valores acima de $0,05$ são satisfatórios para a aceitação da hipótese nula, o que corresponde a dizer que não existem diferenças estatisticamente significativas entre os dois conjuntos. Por convenção, será utilizado $\mathbf{\Phi}$ para indicar a utilização desse teste.
  O segundo teste utilizado foi o erro absoluto relativo [Frery et al. 2008], que considera uma comparação entre as médias dos dados originais $\mathcal{V}$ e reduzidos $\mathcal{V}'$. Esse erro é dado por

$$\mathcal{R}'_{\Upsilon} = 100\, Max\big\{\forall_i \big| (\overline{\mathcal{V}_i} - \overline{\mathcal{V}'_i})/\overline{\mathcal{V}_i} \big|\big\},$$

  onde $\overline{\mathcal{V}_i}$ e $\overline{\mathcal{V}'_i}$ são as médias dos dados originais e reduzidos, respectivamente. O $\mathcal{R}'_{\Upsilon}$ é calculado para cada sensor i e somente o maior deles, situação onde a técnica é menos eficiente, será usado.

A avaliação da representatividade dos dados considera ainda os cenários apresentados na tabela 1. Essa tabela mostra o tipo de dado utilizado, a quantidade de fenômenos monitorados por cada nó em redes planas ($\gamma = p$), a quantidade de sensores por agrupamento nas redes hierárquicas ($\gamma = s$), além da média μ e desvio padrão σ utilizados na geração dos dados. Nesses cenários, foram avaliadas as técnicas PCA, PCA-robusta e

ICA, com o objetivo de verificar com qual dessas técnicas o algoritmo proposto obtém um melhor desempenho. As avaliações foram feitas variando-se o tamanho dos dados n em $\{256, 512, 1024, 2048\}$ e n' em $\{n/2, \log_2 n\}$. Por fim, utilizou-se $\mathcal{R}'_\Phi$ e $\mathcal{R}'_\Upsilon$ como regras de decisão D'.

Tabela 1. Cenários considerados

Cenário	Dado real?	Dado com ruído?	Valor γ em $\mathcal{V}_{\gamma,n}$		Parâmetros da Geração	
			Plana	*Hierarq.*	μ	σ
I	Não	Não	$p = 5$	$s = 400$	$10, 30, 50, 70, 90$	10%
II	Sim	Não	$p = 5$	$s = 400$	12 médias reais por sensor	10%
III	Não	Sim	$p = 5$	$s = 400$	$10, 30, 50, 70, 90$	10%
IV	Sim	Sim	$p = 5$	$s = 400$	12 médias reais por sensor	10%

A primeira análise considera o $\mathcal{R}'_\Phi$, que representa o teste *ANOVA*. Os resultados apresentados na tabela 2 correspondem ao cenário I. Esses resultados indicam que não existem diferenças significativas entre as variâncias dos conjuntos original e reduzido tanto para PCA quanto para PCA-robusta e ICA. Dessa forma, as decisões D' relacionadas a $\mathcal{R}'_\Phi$ podem ser tomadas sem que sejam comprometidas. Em ambas as topologias de rede avaliadas os resultados de PCA e PCA-robusta foram praticamente idênticos e mais significativos que os observados com ICA. No cenário II os resultados foram muito similares aos observados no cenário I, com um alto índice de significância dos mesmos. Nos cenários III e IV, os resultados observados tiveram índice de significância menor, devido à presença de ruído nos dados, mas os mesmos podem ser considerados satisfatórios, uma vez que em todos os casos os valores foram superiores a $0,05$.

Tabela 2. Cenário I – Análise de variância (p-valor)

Rede plana

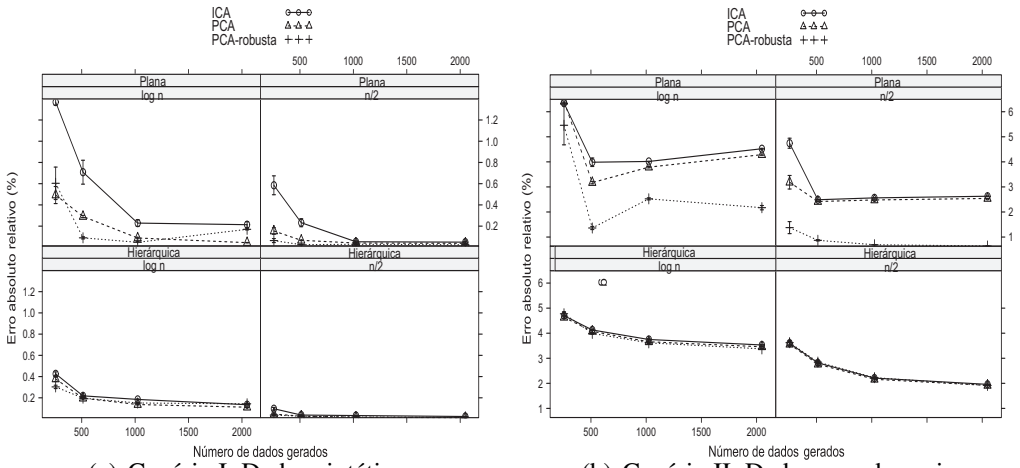
Distribuição	(n = 256)		(n = 512)		(n = 1024)		(n = 2048)	
normal	$n/2$	$\log_2 n$	$n/2$	$\log_2 n$	$n/2$	$\log_2 n$	$n/2$	$\log_2 n$
PCA	0,98	0,86	0,98	0,86	0,98	0,87	0,98	0,87
ICA	0,89	0,83	0,84	0,83	0,76	0,82	0,76	0,83
PCA-robusta	0,98	0,85	0,98	0,86	0,98	0,86	0,98	0,87

Rede hierárquica

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-robusta	0,90	0,81	0,89	0,86	0,89	0,84	0,89	0,82

A segunda análise considera o $\mathcal{R}'_\Upsilon$, que representa o erro absoluto relativo. Os resultados dos cenários I e II podem ser observados na figura 2. Esses cenários consideram, respectivamente, os resultados com dados sintéticos e pseudo-reais sem ruído. Na figura 2(a), considerando a amostragem $n/2$, os resultados obtidos com as três técnicas foram muito satisfatórios, com $\mathcal{R}'_\Upsilon$ próximo a zero em todos os tamanhos de dados gerados. É importante destacar que quando aumentou-se a quantidade de dados gerados, em praticamente todos os casos $\mathcal{R}'_\Upsilon$ diminuiu, o que ocorre porque uma maior quantidade de dados é gerada para cada fenômeno com a mesma média μ e desvio padrão σ. Nesse

cenário, os resultados mais significativos foram observados com as técnicas PCA-robusta e PCA, considerando, respectivamente, as redes planas e hierárquicas, e os menos significativos com a técnica ICA. Isso ocorre pelo fato de que, conforme descrito em [Hyvärinen 1999], a primeira componente independente pode não ter a maior variância dos dados, tornando a amostragem menos eficiente. Especificamente na rede hierárquica os erros são menores que na rede plana, uma vez que todos os sensores no agrupamento monitoram o mesmo fenômeno, o que resulta em maior quantidade de informações relevantes e, consequentemente, em melhor desempenho na amostragem. Com relação à amostragem $\log_2 n$, assim como na amostragem $n/2$, os erros obtidos com as três técnicas foram muito pequenos em ambas as topologias e à medida que aumentou-se a quantidade de dados gerados, $\mathcal{R}'_\Upsilon$ diminuiu, mostrando a escalabilidade do método proposto em relação a esse parâmetro. Com isso, observou-se eficiência do MuSA nesse cenário, mostrando que as decisões D' referentes ao $\mathcal{R}'_\Upsilon$ não seriam comprometidas.



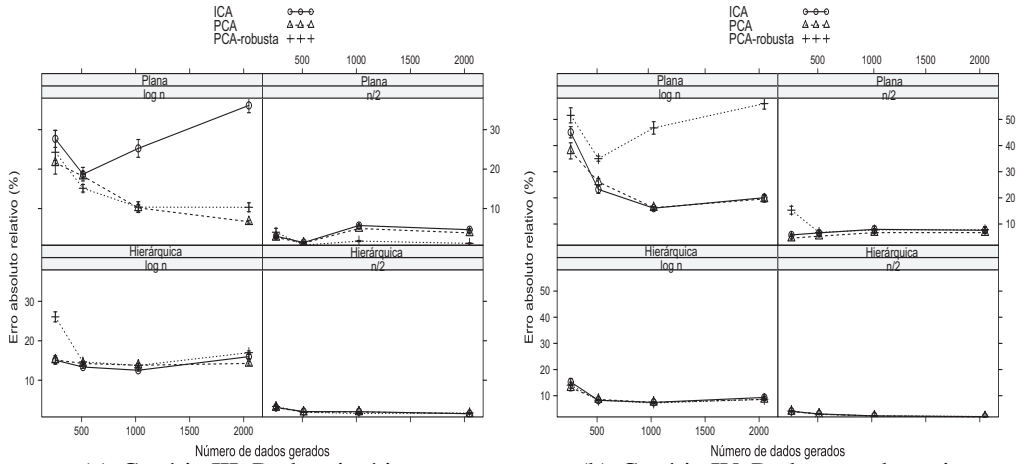
(a) Cenário I: Dados sintéticos (b) Cenário II: Dados pseudo-reais

Figura 2. Análise do MuSA considerando dados sem ruído

Na figura 2(b), considerando a amostragem $n/2$, como na avaliação com dados sintéticos, MuSA obteve resultados satisfatórios. Novamente os melhores resultados foram observados com PCA-robusta e PCA nas redes plana e hierárquica, respectivamente. Mais uma vez pode-se observar a escalabilidade do algoritmo em termos da quantidade de dados sensoriados, com o $\mathcal{R}'_\Upsilon$ diminuindo ou se mantendo praticamente constante quando aumentou-se o número de dados do conjunto original $\mathcal{V}$. Nas simulações com amostragem $\log_2 n$, os $\mathcal{R}'_\Upsilon$ observados foram baixos, sendo próximos aos observados com a amostragem $n/2$, ratificando a viabilidade do uso do algoritmo MuSA para efetuar a amostragem. Nesse caso, os melhores resultados foram obtidos com PCA-robusta em redes planas e hierárquicas. Assim como com dados sintéticos, as decisões D' seriam tomadas sem comprometimento.

Por fim, os resultados das avaliações com ruído, considerando o $\mathcal{R}'_\Upsilon$, podem ser observados na figura 3, que ilustra os resultados dos cenários III e IV. Na figura 3(a), considerando a amostragem $n/2$, mesmo com a presença de ruído nos dados, os resultados foram satisfatórios. Nesse caso, os mais significativos foram observados com PCA-robusta e ICA em redes planas e hierárquicas, respectivamente. Com a amostragem $\log_2 n$, os erros foram muito maiores que na amostragem $n/2$, especialmente com a

amostragem baseada em PCA-robusta. Com essa técnica, quando aumentou-se a quantidade de dados gerados, o $\mathcal{R}'_{\Upsilon}$ também aumentou em quase todos os casos. Entretanto, com PCA e ICA o MuSA foi mais eficiente e os erros diminuíram à medida que aumentou-se a quantidade de dados gerados. Considerando que a variação dos dados originais é significativa devido à introdução de ruído nos dados, os resultados observados podem ser considerados satisfatórios, principalmente para grandes quantidades de dados gerados.



(a) Cenário III: Dados sintéticos (b) Cenário IV: Dados pseudo-reais

Figura 3. Análise do MuSA considerando dados com ruído

Na figura 3(b), com a amostragem $n/2$, mais uma vez os resultados foram satisfatórios. Nesse caso, os mais significativos foram obtidos com PCA em redes planas e hierárquicas, embora as diferenças entre as técnicas sejam pequenas em praticamente todos os tamanhos de dados gerados. Com a amostragem $\log_2 n$, como no cenário III, os erros foram muito maiores que na amostragem $n/2$, sendo os maiores valores observados com a técnica ICA em redes planas. Contudo, utilizando PCA e PCA-robusta, MuSA foi mais eficiente e $\mathcal{R}'_{\Upsilon}$ diminuiu em quase todos os casos quando aumentou-se a quantidade de dados gerados. Com isso, os resultados nesse cenário podem ser considerados satisfatórios, principalmente para grandes quantidades de dados em redes planas e para todos os tamanhos de dados em redes hierárquicas.

Com base nos resultados apresentados, a tabela 3 apresenta as técnicas que obtiveram melhores resultados em cada cenário. Devido à similaridade dos resultados na avaliação do $\mathcal{R}'_{\Phi}$, serão considerados apenas os resultados do $\mathcal{R}'_{\Upsilon}$. A partir desses resultados, é possível observar que PCA aparece em todos os cenários. Além disso, mesmo nos casos nos quais os resultados de PCA não foram os melhores, eles foram próximos e em nenhum caso seu comportamento se mostrou irregular, como ocorreu com PCA-robusta e ICA. Dessa forma, é possível concluir que a amostragem baseada em PCA se mostrou mais eficiente devido a sua regularidade, considerando os cenários avaliados. É importante enfatizar, no entanto, que em ICA a primeira componente pode não ter a maior porcentagem da variância dos dados. Nesse caso, uma ordenação das componentes independentes pode resultar em uma amostragem mais eficiente.

Tabela 3. Melhores resultados para cada cenário

Cenário	Rede plana		Rede Hierárquica	
	$n/2$	$\log_2 n$	$n/2$	$\log_2 n$
I	PCA-robusta	PCA	PCA	PCA
II	PCA-robusta	PCA-robusta	PCA	PCA-robusta
III	PCA-robusta	PCA	ICA	ICA
IV	PCA	PCA	PCA	PCA-robusta

6. Avaliação do comportamento da rede

Esta seção descreve a avaliação do comportamento da rede em termos do consumo de energia e atraso na entrega dos dados, ao efetuar a amostragem através do MuSA. Para essa avaliação, as simulações são feitas usando somente o algoritmo baseado em PCA, porque o processamento realizado pela técnica não afeta o desempenho do MuSA. No simulador utilizado, o modelo de dissipação de energia é linear, com valores pré-estabelecidos para transmissão e recepção, sem considerar o processamento dos dados.

Esta avaliação é feita através do *Network Simulator 2 (NS-2)* versão 2.33 (`http://nsnam.isi.edu/nsnam/index.php/Main\_Page`). Executou-se as simulações com 33 topologias aleatórias e os resultados são apresentados com um intervalo de confiança assintótico simétrico de 95%. As redes usam um algoritmo de roteamento baseado em árvore de menor caminho e todos os nós têm a mesma configuração de hardware. A densidade da rede é mantida constante e os nós fontes são distribuídos aleatoriamente na região sensoriada. Com o objetivo de avaliar somente o desempenho na amostragem, as árvores são construídas apenas uma vez antes que o tráfego se inicie. Alguns parâmetros utilizados nas simulações são apresentados na tabela 4.

Tabela 4. Parâmetros de simulação

Parâmetros	Valores	Parâmetros	Valores
Tamanho da rede	Varia com a densidade	Localização da fonte	Aleatória
Localização do sink	0, 0	Início do tráfego (s)	500
Tempo de simulação (s)	1100	Término do tráfego (s)	600
Alcance do rádio (m)	50	Taxa de dados (s)	60
Largura de banda (kbps)	250	Energia inicial (J)	100

O tamanho da rede varia, mas a densidade é mantida constante em $8,48$ e é dada por $net_t = \sqrt{\pi\, a_r^2\, |S|/8,4791}$, onde a_r é o alcance do rádio e S o número de nós sensores. O tamanho da fila suportada por cada nó varia com o tamanho dos dados para que não haja descarte de pacotes, ou seja, cada nó suporta o tamanho da fila usada pela aplicação. O alcance do rádio e a largura de banda consideram a especificação do sensor MicaZ (`http://www.xbow.com/`). O tempo de simulação foi fixado em 1100 s, onde os 500 s iniciais são usados para construção e configuração da rede e da estrutura de roteamento e os 500 s finais são usados para permitir que os pacotes restantes na rede sejam transmitidos. Com isso, o tráfego real na rede dura 100 s. Além disso, a energia inicial utilizada foi 100 J, para que as reservas de energia dos nós nunca se esgotem.

A avaliação do comportamento da rede considera os seguintes cenários: avaliação do consumo de energia e avaliação do atraso na rede. Considerou-se nesses cenários uma

distribuição normal multivariada sem ruído, com $\mu = 50$ e $\sigma = 10\%$ e os valores γ em $\mathcal{V}_{\gamma,n}$ são $p = 5$ na rede plana e $s = \{50, 100, 150, 200\}$ correspondente ao número de nós por agrupamento, na rede hierárquica. Em todas as avaliações utilizou-se $n' = \{n/2, \log_2 n\}$.

A primeira análise considera o consumo de energia em redes planas. Resultados de simulações são apresentados na figura 4. Nesse caso, avaliou-se as seguintes situações: (i) variou-se o tamanho dos dados gerados em $n = \{256, 512, 1024, 2048\}$, com número fixo de sensores por nó $s = 5$, 5 nós fontes e 128 nós na rede; (ii) variou-se o número de nós em $\{128, 256, 512, 1024\}$, fixando o número de sensores em $s = 5$, a quantidade de dados gerados em $n = 256$ e o número de nós fontes em 5; (iii) variou-se o número de nós fontes em $\{1, 5, 10, 20\}$, fixando o número de sensores por nó em $s = 5$, a quantidade de dados gerados em $n = 256$ e o número de nós em 128. Em todas as situações avaliadas, foi possível observar que a utilização do algoritmo MuSA resultou em considerável redução do consumo de energia na rede, comprovando a viabilidade de utilização da técnica para aumentar o tempo de vida da rede.

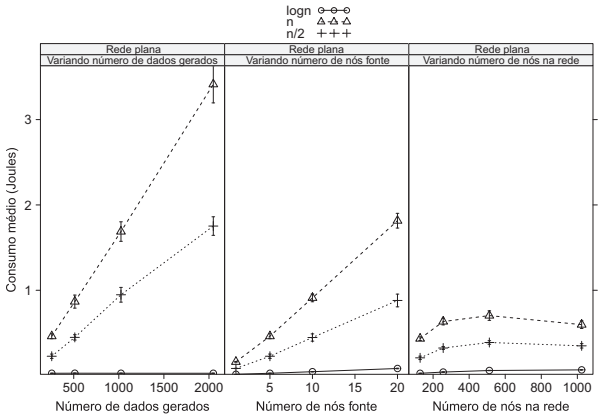


Figura 4. Avaliação do consumo de energia médio na rede

Com relação à rede hierárquica, avaliou-se as seguintes situações: (i) variou-se o tamanho dos dados enviados pelos nós do agrupamento ao nó líder em $n = \{32, 64, 128, 256\}$; (ii) variou-se o número de nós por agrupamento em $s = \{50, 100, 150, 200\}$; (iii) variou-se o número de agrupamentos em $\{2, 4, 6, 8\}$. Em todas as situações simuladas, assim como no caso da rede plana, obteve-se considerável redução no consumo de energia da rede.

A segunda análise do comportamento da rede se refere ao atraso na entrega dos pacotes e segue as mesmas considerações descritas na avaliação do consumo de energia. Nesse caso, o atraso é determinado pelo intervalo de tempo entre o envio e a recepção completa dos pacotes. Os resultados observados com a rede plana são apresentados na figura 5. Assim como ocorreu com o consumo de energia, o algoritmo MuSA reduziu significativamente o atraso na entrega dos pacotes na rede. O mesmo foi observado nas avaliações com a rede hierárquica, ratificando a eficiência e a viabilidade do uso da técnica proposta no que se refere ao comportamento da rede, também para redução do atraso.

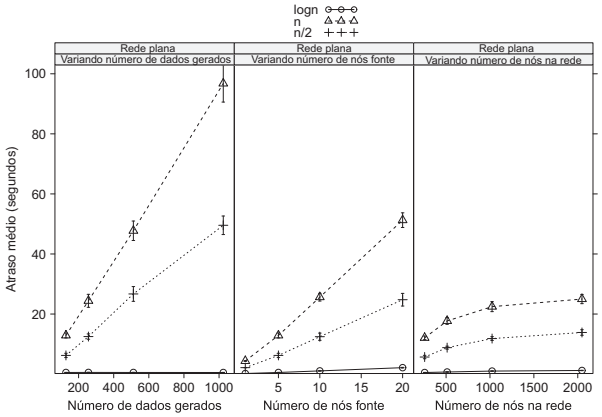


Figura 5. Avaliação do atraso médio na rede

7. Conclusão e trabalhos futuros

RSSFs têm restrições de energia e a extensão de seu tempo de vida é um dos principais problemas no projeto dessas redes. Neste trabalho foi apresentado um algoritmo de amostragem para redução de dados multivariados. MuSA utiliza técnicas de análise de componentes para classificar os dados, permitindo a seleção de uma amostra contendo apenas os dados mais relevantes para a aplicação.

Resultados mostraram a eficiência da técnica em redes planas e hierárquicas. No que se refere à representatividade dos dados, em todos os cenários, o MuSA obteve valores satisfatórios para a análise de variância e erro absoluto relativo, inclusive na avaliação com ruído. Além disso, o método se mostrou escalável em termos da quantidade de dados sensoriados, uma vez que aumentando essa quantidade, os erros se mantiveram aceitáveis e, em muitos casos, diminuíram. Nas redes hierárquicas, os resultados foram ainda mais significativos que nas redes planas. Comparando as técnicas de análise de componentes, PCA foi mais eficiente em grande parte das avaliações e teve comportamento mais regular. Por sua vez, a amostragem baseada em ICA foi menos eficiente na maioria dos casos. Entretanto, é possível melhorar seu desempenho ordenando suas componentes.

O algoritmo MuSA obteve bons resultados também na avaliação do comportamento da rede. Tanto em redes planas quanto em redes hierárquicas, sua utilização resultou em considerável diminuição do consumo de energia na rede. Além disso, o atraso no envio dos pacotes também foi reduzido consideravelmente em ambos os casos, comprovando a viabilidade do uso do MuSA para efetuar a amostragem multivariada em RSSFs.

Como trabalhos futuros, pretende-se efetuar a amostragem juntamente com a tarefa de roteamento, visando obter maior eficiência de energia. Pretende-se também avaliar o MuSA após a ordenação das componentes independentes, para verificar o desempenho da técnica ICA. Outra intenção é de avaliar o MuSA em cenários nos quais perdas dos dados podem afetar sua qualidade e compará-lo com técnicas não baseadas em análise de componentes. Pretende-se ainda realizar as avaliações utilizando simuladores específicos para RSSF e realizar experimentos em plataformas reais de sensores.

Referências

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

Aquino, A. L. L., Figueiredo, C. M. S., Nakamura, E. F., Buriol, L. S., Loureiro, A. A. F., Fernandes, A. O., and Junior, C. N. C. (2007). Data stream based algorithms for wireless sensor network applications. In *Proceedings of 21st IEEE International Conference on Advanced Information Networking and Applications (AINA'07)*, pages 869–876, Niagara Falls, Canada. IEEE Computer Society Press.

Cvejic, N., Bull, D., and Canagarajah, N. (2007). Improving fusion of surveillance images in sensor networks using independent component analysis. *IEEE Trans. Consum. Electron.*, 53(3):1029–1035.

Frery, A. C., Ramos, H., Alencar-Neto, J., and Nakamura, E. F. (2008). Error estimation in wireless sensor networks. In *Proceedings of 23rd ACM Symposium on Applied Computing 2008 (SAC'08)*, pages 1923–1927, Fortaleza, Brazil. ACM.

Hyvärinen, A. (1999). Survey on independent component analysis. *Neural Comput. Surv.*, 2(1):94–128.

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

Li, J. and Zhang, Y. (2006). Interactive sensor network data retrieval and management using principal components analysis transform. *Smart Mater. Struct.*, 15(11):1747–1757.

Nakamura, E. F., Loureiro, A. A. F., and Frery, A. C. (2007). Information fusion for wireless sensor networks: Methods, models, and classifications. *ACM Comput. Surv.*, 39(3):1–55.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philos. Mag.*, 2(6):559–572.

Roy, O. and Vetterli, M. (2008). Dimensionality reduction for distributed estimation in the infinite dimensional regime. *IEEE Trans. Inf. Theory*, 54(4):1655–1669.

Ruymgaart, F. H. (1981). A robust principal component analysis. *J. Multivariate Anal.*, 11(4):485–497.

Santini, S. and Romer, K. (2006). An adaptive strategy for quality-based data reduction in wireless sensor networks. In *Proceedings of 3rd International Conference on Networked Sensing Systems (INSS'06)*, pages 29–36, Chicago, IL, USA.

Seo, S., Kang, J., and Ryu, K. H. (2005). Multivariate stream data reduction in sensor network applications. In *Proceedings of 2nd International Symposium on Ubiquitous Intelligence and Smart Worlds (UISW'05)*, pages 198–207, Nagasaki, Japan. Springer.

Thomson, N. (1993). Understanding ANOVA the APL way. *ACM SIGAPL – APL Quote Quad*, 24(1):295–303.

Zhu, J. and Papavassiliou, S. (2004). A resource adaptive information gathering approach in sensor networks. In *Proceedings of IEEE Sarnoff Symposium on Advances in Wired and Wireless Communication (SARNOFF'04)*, pages 115–118, Princeton, USA. IEEE Computer Society.