

# Uma Análise da Distorção na Coleta de Dados em Agrupamentos Baseados em Similaridade Espacial

Sérgio Samuel Furlaneto <sup>1</sup>, Michel Albonico <sup>2</sup>

<sup>1</sup>Universidade Federal do Paraná (UFPR)  
Caixa Postal 81530-900 – Curitiba – PR – Brazil

sersamfur@gmail.com

<sup>2</sup>Departamento de Sistemas e Computação  
Universidade Regional de Blumenau (FURB) – Blumenau, SC – Brazil

michel@inf.ufrgs.br

## 1. Introdução

A coleta dos dados em redes de sensores sem fio (RSSF) de médio e grande porte é uma tarefa muito difícil, visto que a troca de mensagens é a tarefa de maior consumo de energia e ocupação de canal de comunicação. A disseminação dos dados pode ser feita de maneira programada, contínua e sob demanda. A modalidade de disseminação sob-demanda é mais eficiente por retornar apenas as informações desejadas pelo usuário, prolongando assim a vida útil da rede [Fung et al. 2002]. Porém, a coleta de dados sob demanda apenas não é suficiente para prover escalabilidade às RSSF.

Dentre as técnicas que diminuem o fluxo de mensagens e proveem escalabilidade dentro das RSSF podemos citar a de agrupamentos. Os agrupamentos de sensores consistem em sub-conjuntos de sensores associados segundo critérios de proximidade e valores de atributos [Bo et al. 2008]. O critério de proximidade consiste em uma distância física fixa, medida em saltos, para delimitar o tamanho máximo do agrupamento. Os atributos de um sensor descrevem suas propriedades e podem ser categorizados como *atributos inerentes ao dispositivo* e *atributos de dados sensorizados*. Os atributos inerentes a sensores são a carga do dispositivo, localização geográfica, altitude e mobilidade. Alguns exemplos de atributos de dados sensorizados são valores de temperatura, pressão, umidade e luminosidade, entre outros.

Em alguns ambientes de sensoriamento, como o urbano, os dados lidos apresentam similaridade espacial. Nestes casos, a formação de agrupamentos baseada nesta característica representa uma maneira mais econômica e mais eficiente de coletar os dados da rede. A definição de agrupamentos baseada em similaridade espacial dos dados, considera explicitamente os atributos dos dispositivos (nós) além de sua localização geográfica. Estes agrupamentos resultam de particionamentos com três propriedades [Reis et al. 2007]:

1. **Homogeneidade interna do agrupamento:** Dispositivos pertencentes ao mesmo agrupamento tem que ser similares aos demais segundo algum atributo pré-definido.
2. **Heterogeneidade entre agrupamentos:** Dispositivos pertencentes a agrupamentos diferentes devem ser diferentes uns dos outros segundo algum atributo pré-definido.

3. **Proximidade ou contiguidade:** Dispositivos pertencentes a um mesmo agrupamento devem pertencer a uma estrutura de vizinhança pré-definida.

A coleta dos dados em agrupamentos formados a partir da similaridade espacial oferece uma maior escalabilidade e disponibilidade dos dados. Porém, a escolha de apenas um valor para representar as leituras de todos os nós do agrupamento pode comprometer a precisão. Sempre que uma mensagem é entregue na estação base com uma variação em relação ao valor real em campo é dito que esta leitura sofreu uma distorção.

As distorções dos dados decorrem da disponibilidade finita de dispositivos para sensoriamento do campo. As distorções sofridas pelos dados podem ser classificadas como espaciais e temporais. O uso da leitura do sensor mais próximo de determinada coordenada ou o uso do valor médio das leituras coletadas em um agrupamento como resposta a uma consulta, em geral, resulta em um erro de precisão dos dados. Este tipo de erro é classificado como uma distorção espacial dos dados. A distorção temporal dos dados consiste na variação sofrida pelos dados no ambiente após sua coleta até sua entrega na estação base. O número de saltos exigido para a disseminação e coleta de dados para uma consulta determina a probabilidade de que os dados estejam defasados [Zhang et al. 2008].

Em [Cristescu and Vetterli 2005] é feita uma análise do comportamento da distorção espaço-temporal dos dados visando determinar a densidade ideal para uma RSSF. O trabalho [Zhang et al. 2007] avalia a distorção espaço temporal dos dados levando em conta critérios de congestionamento da rede e escalonamento de tarefas da rede. A precisão dos sumários de dados coletados dos líderes de agrupamentos é analisado em [Reis et al. 2007], porém seu estudo se restringe ao estudo da representatividade dos dados coletados em um agrupamento baseado em similaridade espacial.

O objetivo principal deste estudo é analisar a distorção espacial dos dados em agrupamentos baseados em similaridade espacial. E mostrar que o uso de agrupamentos favorece uma maior disponibilidade dos dados e o uso da similaridade espacial dos dados em sua formação atenua a distorção dos dados.

## 2. Trabalhos Relacionados

Nesta seção apresentaremos alguns trabalhos que estudam o fenômeno da distorção da distorção dos dados e que propõe algum tipo de solução para este problema.

### Optimal Clustering

A distorção dos dados é afetada diretamente pela densidade das redes de sensores. Sendo assim, o trabalho [Cristescu and Vetterli 2005] estuda a distorção sofrida pelos dados com o aumento e diminuição da densidade da rede, visando obter uma densidade da rede que proporcione uma menor latência com a maior precisão possível.

Neste trabalho o cálculo do tempo necessário para a transmissão de pacotes é expresso de maneira simplificada como  $kt$ , onde  $k$  representa o número mínimo de saltos exigidos até a entrega de uma mensagem e  $t$  o tempo necessário necessário para o envio de cada pacote. No mundo real, o atraso depende de diversos fatores, dos quais podemos citar o congestionamento da rede e perda de pacotes, entre outros.

## Coleta de Dados em Tempo Real

O problema da distorção espaço-temporal dos dados é estudada por [Zhang et al. 2007]. As principais preocupações neste trabalho são com o consumo de energia e precisão dos dados coletados.

Devido ao maior número de fatores que podem afetar a transmissão de pacotes nas RSSF, é proposto um arcabouço para cálculo da distorção espaço-temporal levando em conta a existência de perdas durante a comunicação decorrente de congestionamento.

Este modelo leva é direcionado a redes heterogêneas, nas quais os *cluster-heads* possuem energia muito mais abundante que os demais integrantes de cada agrupamento. A partir desta suposição, o arcabouço é alimentado e provê uma relação mais eficiente entre as dimensões da rede e o número de *cluster-heads* nela dispersos.

Neste trabalho a distorção é calculada sem ser levado em conta a existência de mecanismos para atenuá-la, como alguns propostos especificamente para campos de sensoriamento que apresentam similaridade espacial dos dados [Reis et al. 2007] ou o uso de operadores de fusão de dados para diminuir a troca de mensagens e então a latência na entrega dos sumários de dados [Nakamura and Loureiro 2008].

## SKATER

Um protocolo de construção de agrupamentos para RSSF's baseada na geografia é proposto no trabalho "*Data-Aware Clustering for Geosensor Networks Data Collection*" (SKATER) [Reis et al. 2007]. Ele considera a similaridade espacial dos dados em adição à sua localização. Sua hipótese é que os sumários de dados provenientes de agrupamentos que apresentam similaridade espacial dos dados são mais precisos e portanto menos distorcidos que os sumários provenientes de agrupamentos que levam em conta apenas a distância do *cluster-head* [Bo et al. 2008].

O foco deste trabalho são aplicações que requerem a coleta contínua de dados e que aceitam sumários dos dados de sub-regiões do campo de sensoriamento. Estes sumários de dados podem se referir aos obtidos através de operadores de média (AVG), que consistem na soma de todos os valores coletados dos sensores de cada agrupamento e divisão por seu número de elementos. Esta técnica também propõe o uso do valor médio do agrupamento caso a leitura de algum dispositivo não possa ter sido coletada.

Este trabalho não trata a distorção espacial e temporal sofrida pelos dados nos sensores que os coletaram. Seu estudo se restringe à distorção causada pelo uso de sumários provenientes de agrupamentos que exploram ou não a similaridade espacial dos dados.

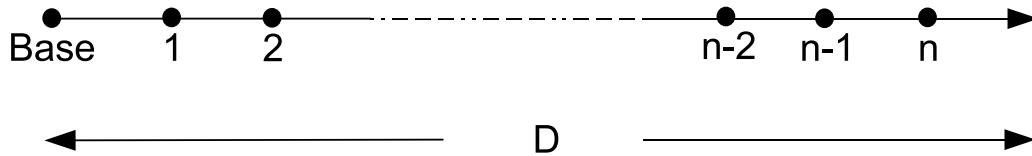
## 3. Modelo de Distorção

Esta seção formaliza o modelo de distorção espacial empregado na análise de precisão dos dados e latência de consultas pontuais e espaciais. As consultas pontuais serão simuladas em um ambiente unidimensional para facilitar a sua representação. Um ambiente bidimensional será usado para avaliação das consultas espaciais.

### 3.1. Ambiente Unidimensional

É assumido que todos os dispositivos sensores da rede (nós) estão distribuídos de maneira homogênea. Após a rede ser formada é escolhido um caminho  $P$ , em linha, arbitrário com

$n$  nós, como ilustra a Figura 1. Nela, a estação base é posicionada no início da linha. A distância  $d$  entre dois dispositivos quaisquer sobre esta linha, em hops, é calculada como  $d = |p_s - p_d|$ , onde  $p_s$  é a posição do sensor emissor da mensagem e  $p_d$  a posição do dispositivo destino.

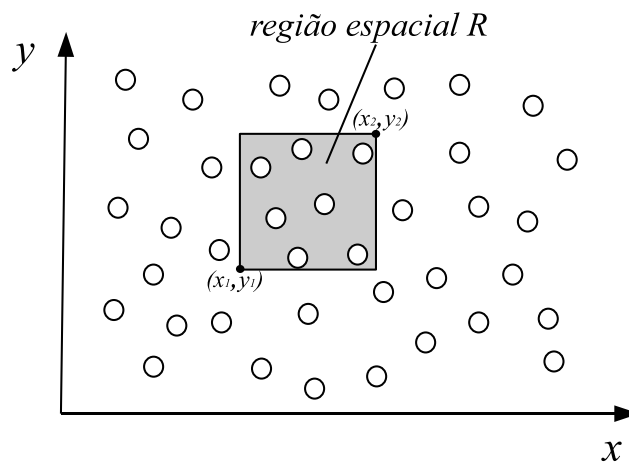


**Figura 1. Ambiente unidimensional**

Seja um caminho  $P$  composto por  $n$  sensores dentro do grafo da rede. Este caminho consiste em uma série de dispositivos que compoem uma semi-reta diagonal a toda a rede. As consultas pontuais consistem na requisição do valor de determinado atributo em uma coordenada específica, onde  $(x_1, 0)$  é a coordenada alvo da consulta. O dispositivo mais próximo da coordenada  $(x_1, 0)$  responde a consulta.

### 3.2. Ambiente Bidimensional

O ambiente de simulação para as consultas espaciais consiste de um grafo bidirecionado, cujos nós são dispersos aleatoriamente como ilustrado na Figura 2.



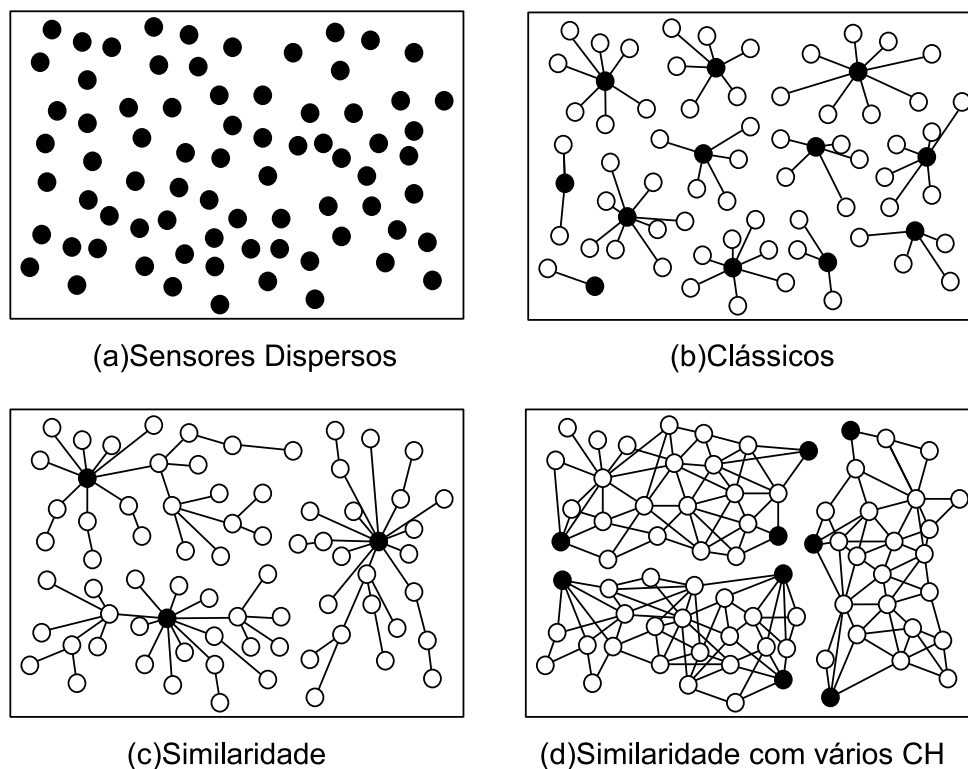
**Figura 2. Ambiente bidimensional**

As consultas espaciais disseminadas sobre este ambiente consistem de duas coordenadas que delimitam uma região espacial  $R$  como ilustra a Figura 2. As consultas espaciais sobre uma região  $R$  são delimitadas por coordenadas  $(x_1, y_1)$  e  $(x_2, y_2)$ .

### 3.3. Modelos de Rede

Neste trabalho, a análise da distorção espacial e latência na obtenção dos dados serão medidas durante a disseminação de consultas pontuais e espaciais em quatro ambientes. A Figura 3.3 ilustra as redes que terão a precisão e latência. A Figura 3.3(a) mostra uma rede não estruturada, 3.3(b) ilustra uma rede de sensores cujos agrupamentos são formados a partir da distância máxima do *cluster-head* (CH), agrupamentos formados a partir

da similaridade espacial dos dados são mostrados em 3.3(c) e em 3.3(d) são mostrados agrupamentos formados a partir da similaridade espacial dos dados, porém com vários *cluster-heads* posicionados nos pontos de encontro das bordas de cada agrupamento.



**Figura 3. Modelos de rede**

Neste trabalho as redes planas são utilizadas como referência para o cálculo da distorção dos dados. Nestas redes as requisições de dados são respondidas apenas pelo nó mais próximo da coordenada desejada.

A avaliação será feita através do uso de métricas de latência média e precisão dos dados. A métrica de latência média representa o tempo médio de resposta de consultas. Nos ambientes em que a coleta é feita em tempo real é vital que o tempo de resposta de qualquer consulta seja próxima a zero, pois os dados podem não serem mais representativos [Zhang et al. 2007]. A latência média é calculada através da soma da latência de todas as consultas disseminadas e então cálculo de sua média aritmética. A precisão dos dados consiste na diferença no valor das leituras entregues à estação base e às leituras presentes no campo. O cálculo da precisão média dos dados é calculado através da soma da diferença do valor das leituras entregues e aos dados do campo [Zhang et al. 2008].

Serão avaliadas as distorções sofridas pelos dados em redes planas, e hierárquicas com um ou mais líderes de agrupamentos (*cluster-heads*). Estas comparações servem para determinar quais destas redes são mais eficientes relacionando o tempo de resposta e a precisão dos dados em cada uma.

A Tabela 1 ilustra estes parâmetros de simulação para a avaliação dos cálculos teóricos e comparação com demais trabalhos relacionados serão feitas simulação na ferramenta ns-2. O ambiente alvo possui dimensões de 200x200 metros para a simulação de

Densidade	Dimensões (m)	Nºde Sensores	Sensor/m <sup>2</sup>
Esparsa	400x400	1000	0,00625
Densa	200x200	1000	0,025

**Tabela 1. Parâmetros de simulação**

redes densas e 400x400 metros para redes esparsas. Em cada rede serão dispersos 1000 dispositivos sensores.

#### 4. Cronograma e Atividades

Mês	Novembro				Dezembro			
Semana	1	2	3	3	1	2	3	4
Estudar funcionamento do ns-2	X	X						
Implementar ambiente de simulação unidimensional		X	X					
Simular disseminação de consultas			X					
Implementar ambiente de simulação bidimensional				X				
Simular disseminação de consultas				X	X			

**Tabela 2. Cronograma**

#### Referências

- Bo, W., Han-ying, H., and Wen, F. (2008). An improved leach protocol for data gathering and aggregation in wireless sensor networks. In *ICCEE*, pages 398–401, Washington, DC, USA. IEEE Computer Society.
- Cristescu, R. and Vetterli, M. (2005). On the optimal density for real-time data gathering of spatio-temporal processes in sensor networks. In *IPSN*, page 21, Piscataway, NJ, USA. IEEE Press.
- Fung, W. F., Sun, D., and Gehrke, J. (2002). Cougar: the network is the database. In *SIGMOD*, pages 621–621, New York, NY, USA. ACM.
- Nakamura, E. F. and Loureiro, A. A. F. (2008). Information fusion in wireless sensor networks. In *SIGMOD*, pages 1365–1372, New York, NY, USA. ACM.
- Reis, I. A., Câmara, G., Assunção, R., and Monteiro, A. M. V. (2007). Data-aware clustering for geosensor networks data collection. In *SBSR*, pages 6059 – 6066, São José dos Campos. INPE.
- Zhang, X., Wang, H., and Khokhar, A. (2008). Distortion analysis for real-time reconstruction of correlated data field in heterogeneous sensor networks. In *GLOBECOM*, pages 1–5.
- Zhang, X., Wang, H., and Khokhar, A. A. (2007). Distortion analysis for real-time data gathering of spatially-temporally correlated fields in sensor networks. In *GLOBECOM*, pages 828–832.