Anais XIII Simpósio Brasileiro de Sensoriamento Remoto, Florianópolis, Brasil, 21-26 abril 2007, INPE, p. 6059-6066.

Data-Aware Clustering for Geosensor Networks Data Collection

Ilka Afonso Reis [1,2]
Gilberto Câmara [1]
Renato Assunção [2]
Antônio Miguel Vieira Monteiro [1]

[1] Divisão de Processamento de Imagens - Instituto Nacional de Pesquisas Espaciais - INPE
Av. dos Astronautas, 1758 - 12227-010 São José dos Campos – SP, Brazil
{ilka,gilberto,miguel}@dpi.inpe.br

[2] Departamento de Estatística - Universidade Federal de Minas Gerais - UFMG
Campus Pampulha - 31270-901 Belo Horizonte – MG, Brasil
assuncao@est.ufmg.br

Abstract. Geosensor networks comprise small electro-mechanical devices that communicate over a wireless network. These devices collect environmental measures and send them to a base station. Energy consumption and data routing are critical factors for efficient geosensor networks. The usual cluster-based data routing protocols for sensor networks group the nodes based on their geographical closeness and aggregate their data to save energy. However, this clustering procedure does not produce the best data summaries. We propose to group the nodes into spatially homogeneous clusters, which consider both the geographical distance and the similarity of measurements between the nodes. Through simulated experiments, we have concluded that spatially homogeneous clusters produce better spatial zones identification and data summaries with a higher statistical quality if compared with the usual clustering methods. Besides, spatially homogeneous clustering can be seen as a tool for spatial sensor data mining, since their clusters represent the partition of the sensor field that has maximum internal homogeneity regarding the values of monitored variable. To make possible the use of our data-aware clustering proposal to collect the sensors' data, we present a design guideline for a cluster-based data routing protocol, the HR-DASH.

Palavras-chave: remote sensing, environment monitoring, spatio-temporal data, data aggregation.

1. Introduction

The advances in wireless and miniaturisation technologies are making possible the development of the *sensor networks*, a new instrument for the remote sensing of the physical world (Elson and Estrin, 2004).

Sensor networks are composed by a large number of small nodes. These nodes are electro-mechanical devices that measure environmental characteristics such as temperature, pressure, humidity and luminosity (**Figure1**, left side). These data are disseminated through wireless communication among the nodes until a base station is reached. Once sensor networks are deployed in the study region, they work without human attendance.

The environmental monitoring is one of many potential uses of this emerging technology (Xu, 2002), especially for hostile environments. According to Martinez et al. (2004), the sensor networks will make possible a realistic monitoring of the natural environmental. In their work, the authors discuss how the environmental monitoring evolved from data logging to sensor networks and describe the GlacsWeb project, an ongoing research in subglacial bed deformation.

Other environmental applications involving sensor networks are described in the literature. Among them, we have the monitoring of the environment of rare and endangered species of plants in a volcano neighboring (Biagioni and Bridges, 2002); the monitoring of the habitat of seabirds (Mainwaring et al., 2002); the microclimate monitoring throughout the volume of giant trees (Culler et al., 2004); the flood monitoring to provide warnings and the monitoring of coastal erosion around small islands (EnviSense-SECOAS). Until recently, experiments have been run on small-scale sensor networks and no large-scale networks have

yet been deployed in practice. However, as the sensors become smaller and cheaper (Warneke et al., 2001), sensor nodes are expected to be densely deployed in the environment.

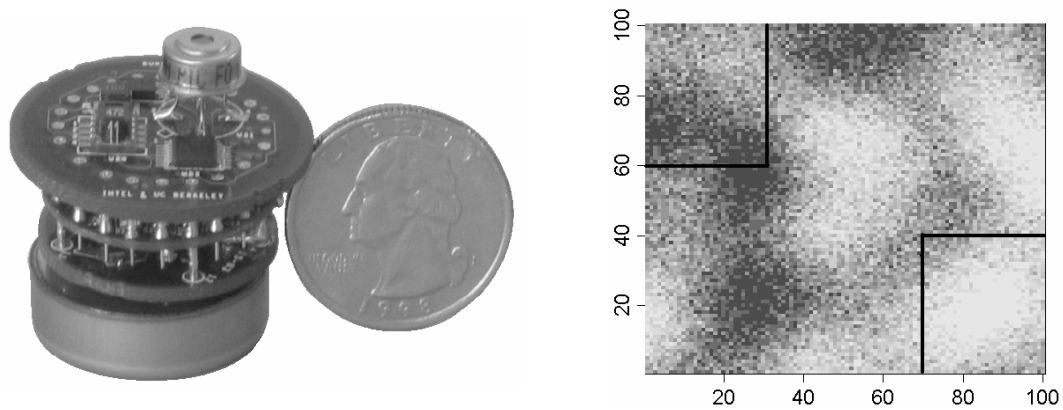

Figure 1. Left side: MICA-2 (by Intel and University of California- Berkeley). Rigth side: spatial distribution of luminosity measurements

Some sensor networks are designed to collect data whose geospatial information is important. To stress their geographic characteristic, these networks are usually defined as *geosensor networks* (Nittel and Stefanidis, 2005). The main goal of a geosensor network is to collect geospatial data while keeping the energy consumption at an acceptable level.

Geosensor networks are an application-driven technology. The temporal resolution of the data determines their delivery model while the required spatial resolution defines the degree of data summarization. Tilak et al. (2002) have identified three data delivery models: continuous data collection and delivery; continuous data collection but the data delivery is triggered by pre-defined events; and on-demand data collection and delivery (queries). On the spatial resolution, some applications need the raw data of all sensing points (Chu et al., 2006; Tulone and Madden, 2006), whereas others need just a summary of all sensors' data, as those TAG (Madden et al., 2002) has been designed for. In the middle of these two extreme cases, there are applications that accept an intermediate degree of data summarization, as maps of temperature and relative humidity, for instance. These applications can have goals as identifying zones of interest such as hot and cold zones. Sensors' data are summarized over subregions, pre-defined (Goldin, 2006) or not, and the spatial distribution of these summaries provides a report of the data variability over the entire region.

In this paper, we are interested in applications that require a continuous data delivery and an intermediate degree of data summarization.

For continuous data delivery, the hierarchical cluster-based data routing protocols are considered to be the most energy efficient alternative (Heinzelman et al., 2002). Multiple cluster-based protocols as LEACH and LEACH-C (Heinzelman et al., 2002) are suitable for applications that admit data summaries over subregions of the sensor field. A cluster-based protocol assembles the sensor nodes into clusters before the data transmission. Except for the clustering procedures as in Tulone and Madden (2006), the usual clustering algorithms considers only the nodes closeness, which we define as *ordinary spatial clustering*. A node chosen as the cluster head receives data from all nodes in its cluster, aggregates these data and sends the summary to the base station. Clustering the nodes keeps most of the communication inside the clusters while data aggregation reduces the messages volume traveling through the network. These strategies together allow for energy saving.

Data aggregation presumes nearby nodes have correlated data. Thus, they are similar to each other and one can aggregate the nodes' data of an ordinary spatial cluster to represent this cluster.

We agree with this reasoning but we believe presuming data correlation is not enough to produce data summaries that are the best estimates of the summarized data. A partition of nodes that considers only their geographical location is missing the most important: the measurements themselves. To make our point clear, consider **Figure 1** (right side), which presents the spatial distribution of luminosity measurements, for instance. Suppose we regularly deploy a geosensor network in the region. The area delimited at right bottom corner has a lower spatial variability in its measurements than the delimited area at upper left corner. Suppose we use one single cluster to summarize the data of each area. The data summary of the first area estimates better the summarized data than the data summary of the second area. Besides, a single cluster could summarize the data in the first area whereas the second area would require a larger number of clusters, to account for the increased spatial variability. To capture these different requirements, the nodes partition might consider the nodes measurements in addition to their geographical location.

Based on these considerations, we present the contributions of this paper.

1.1. Our proposal

We propose a data-aware clustering procedure that groups the nodes considering the *spatial homogeneity* of the nodes' data in addition to their location. Our hypothesis is that data summaries based on *spatially homogeneous clusters* will have a better statistical quality if compared with data summaries based on ordinary spatial clusters. A statistical quality measure expresses how well the data summary sent to the base station estimates the data collected by the nodes.

In this paper, our major concern is to examine how the spatial arrangement of the clusters in a geosensor network affects the statistical quality of the data received by the base station. We compare spatially homogeneous clusters with ordinary spatial clusters regarding the statistical quality of their summaries.

To make possible the use of our clustering proposal in the data collection and address the issue of energy saving, we propose a guideline to design the HR-DASH. The HR-DASH is a cluster-based routing protocol that uses a spatially homogeneous partition of the nodes instead of the usual ordinary spatial clusters.

The remainder of this paper is organized as follows. In section 2, we define the spatially homogeneous clusters and give a brief description of SKATER, the procedure for obtaining such clusters. Section 3 presents the main results of simulated experiments comparing ordinary and spatially homogeneous clusters, based on the statistical quality of their data summaries. In section 4, we present the main features of the HR-DASH protocol. Finally, section 5 draws some concluding remarks.

2. Spatially Homogeneous Clusters

In contrast to ordinary spatial clusters, the definition of spatially homogeneous clusters considers explicitly the nodes' attributes besides their geographical location. Spatially homogeneous clusters are clusters resulting from a partition procedure with three properties.

First, nodes belonging to same cluster have to be similar to each other in some predefined attributes (cluster internal homogeneity). Second, nodes belonging to different clusters have to be different from each other (heterogeneity among clusters). Third, the nodes of a same cluster must belong to a predefined neighborhood structure (closeness or contiguity). The clustering proposed by Tulone and Madden (2006) assembles clusters based on the similarity between the

head node and the nodes inside its range of communication. However, there is no warranty the first and second properties are satisfied. So, they cannot be classified as spatially homogeneous clusters.

To get the spatially homogeneous clusters, we propose the use of the spatial clustering algorithm developed by two of the authors, the SKATER (Spatial 'K'luster Analysis by Tree Edge Removal) (Assunção et al., 2006). This algorithm is a strategy for transforming the regionalisation problem into a graph partitioning problem. SKATER works in two steps. First, it creates a minimal spanning tree (MST) from the graph representation for the neighborhood structure of the geographic entities. The cost of an edge represents the similarity of the entities' attributes, defined as the Euclidean squared distance between them. The MST represents a statistical summary of the neighborhood graph based on the entities' attributes. In the second step, SKATER performs a recursive partitioning of the MST to get contiguous clusters. The MST partitioning considers explicitly the clusters internal homogeneity.
In the geosensor networks context, the graph vertices are the sensor nodes and the cost of an edge connecting a pair of vertices is the similarity between the nodes' data.

Spatially homogeneous clustering offers the possibility of transforming the undelivered raw data into information, since its clusters represent the partition of the sensor field that has maximum internal homogeneity regarding the values of monitored variable. This information cannot be directly extracted from the summaries based on ordinary spatial clusters or on the proposed clusters in Tulone and Madden (2006). As a result, spatially homogeneous clustering can be seen as a tool for spatial sensor data mining.

3. Assessing the Performance of the Spatially Homogeneous Clusters

This section presents the main results of the simulated experiments we have carried out to provide a preliminary evaluation of the performance of the spatially homogeneous clusters. Some results were not presented here for briefness.

3.1. The simulated experiments

We have simulated datasets with spatial autocorrelation using a grid of 10000 cells (100 x 100). We refer these data as *original data*. These datasets were characterized by *extreme zones*, which are clusters with high values (*clear zones*) and clusters with low values (*dark zones*). The zones size was controlled by a scale parameter. **Figure 1** presents the spatial distribution of a dataset simulated using a scale parameter equal to 20. We have also used the values 5, 10 and 15. The higher the scale value, the larger the zone size. To each value of the scale parameter, we have simulated 10 spatial datasets.

We have delimited the extreme zones of the spatial datasets using techniques of image classification. The result was the *zones image* (**Figure 2**, left side), which present the clear and the dark zones. We have used these images to evaluate the ability of the clustering methods to identify the extreme zones.

To sample the spatial datasets, we have deployed a geosensor network with 100 nodes in a regular fashion, as illustrated by the black dots at the right side of the **Figure 2**. The data collected by a sensor is the value of the cell in which it was placed. We refer to this sample as *geosensor data*. The sensor's data was assigned to all cells belonging to its *area of influence*, that is, the set of cells of which the sensor was the nearest node.

To choose the number of clusters to be assembled by the clustering methods, we have adopted the expression proposed by Heinzelman et al. (2002). This expression finds the optimal number of clusters that minimizes the total energy dissipated in a data transmission of the LEACH. Adopting the radio energy model as in Heinzelman et al. (2002), we have found six as the optimal number of clusters.

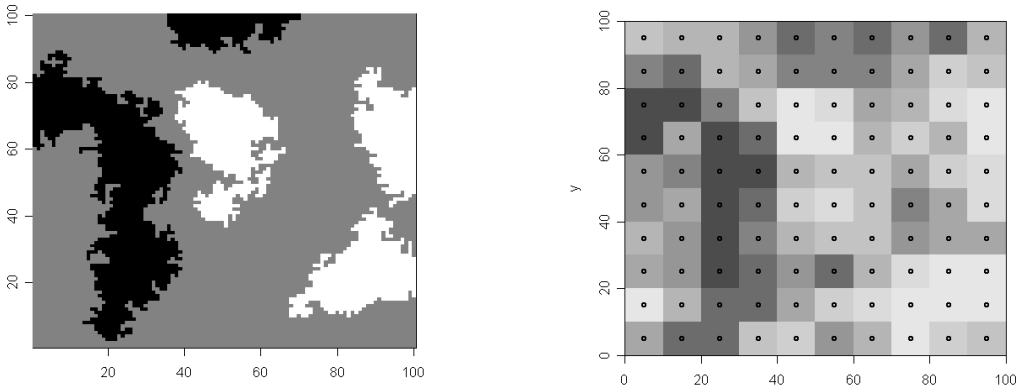

Figure 2. Left side: zones image of the dataset in Figure 1. Right side: geosensor data and the network's nodes.

To get the ordinary spatial clusters, we have simulated the LEACH clustering procedure (Heinzelman et al., 2002). We have chosen the cluster heads randomly among all nodes, but constrained to a minimal distance of 30 meter between them. This constraint tries to simulate the choice of cluster heads by LEACH-C (Heinzelman et al., 2002), avoiding to agglomerate the head nodes. To assemble the clusters, we have associated the remaining nodes to their nearest cluster head. Since the cluster heads location can affect the performance of ordinary clusters, we have used 10 different cluster heads arrangements for each spatial dataset.

The spatially homogeneous clusters were obtained by the SKATER procedure (Assunção et al., 2006).

To each cluster *k*, we have calculated the cluster summary CM_k as the average of the data of the nodes belonging to the cluster *k*. We have defined the *statistical quality of a cluster summary* (SQ_k) using the following expression

$$SQ_k = \sum_{i=1}^{N_k} \frac{1}{N_k} \cdot \frac{\left| x_{ik} - \mathrm{CM}_k \right|}{x_{ik}},$$

where x_{ik} is the original data of the cell *i* that belongs to the union of the area of influence of all nodes of the cluster *k* ; N_k is the number of cells belonging to the area of influence of the cluster *k*. The statistical quality measure SQ_k is the average relative error of the cluster mean CM_k. It measures how far the cluster summary CM_k is from the original data, in average, when these original data are replaced by CM_k. The smaller the values of SQ_k, the better the cluster summary CM_k represents the individual cell values. SQ_k is a measure of the local performance of the clusters summaries.

3.2 The results

We have evaluated the SQ_k values of the spatially homogeneous (SH) and the ordinary spatial (O) partitions. **Figure 3** presents the boxplots[1] for SQ_k values according to the scale parameter.

For spatial datasets with the smallest scale, the ordinary clusters have had a better local performance. The geosensor data collected at these spatial datasets simulate data with no spatial autocorrelation (random data). The geosensor data could not capture the spatial patterns of the

[1] The bottom and the top of the box represent the percentiles 25 and 75, respectively. Therefore, the box height is a measure of the data variability. The line drawn across the box represents the median and the points outside the dashed lines represent the outlier values. If there are not outliers, the ends of the inferior and superior dashed lines represent the minimum and the maximum values, respectively.

original data, because the distance between the nodes was larger than the zone size. The superior performance of the ordinary clusters happens because ordinary clustering groups the nodes in a random fashion, whereas the spatially homogeneous clustering tries to found patterns that geosensor data were not able to capture.

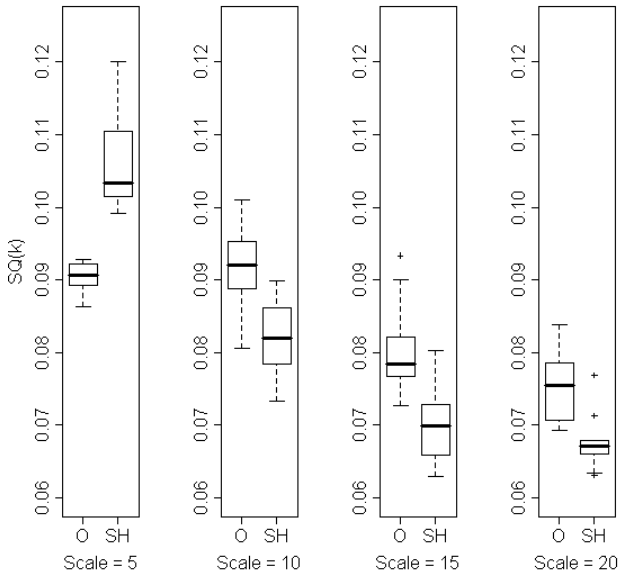

Figure 3. Boxplots for SQ_k values.

At the largest scales (10, 15 and 20), the clusters produced by both clustering methods have a good statistical quality. The SQ_k median values are smaller than 0.10, which is the coefficient of variance of the simulated spatial datasets. The larger the scale parameter, the higher is the statistical quality of the clusters. The most of the SQ_k values of spatially homogeneous clusters are smaller than the values of ordinary clusters. Thus, when geosensor data are able to capture the spatial patterns of the original data, the local performance of the spatially homogeneous clusters is better than the performance of the ordinary clusters.

We have used the summaries CM_k to classify the clusters into clear or dark, using a criterion similar to that used to produce the zones image. **Figure 4** presents the results of this classification for the two clustering methods, using the geosensor data of the **Figure 2**. The black clusters are those classified as dark and the white clusters represent the clear clusters.

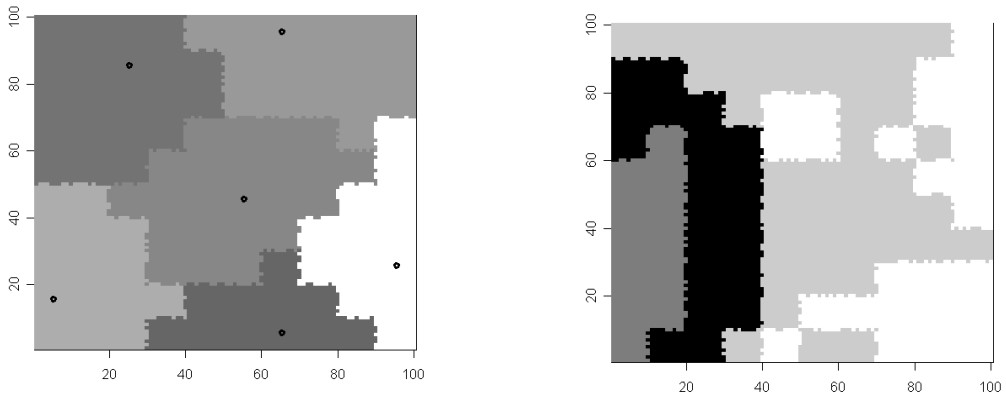

Figure 4. Two partitions of the geosensor network of the Figure 2: ordinary spatial clusters (left side) and spatially homogeneous clusters (right side).

Comparing **figures 2** and **4**, we see the spatially homogeneous clusters could identify more extreme spatial zones than the ordinary spatial clusters.

The comparison between **figures 2** and **4** represents what happens with the spatial datasets simulated using the other scale parameters. The spatially homogeneous clusters was able to identify zones even when they were small and the geosensor data did not seem to reveal any aspects of spatial autocorrelation (scale parameter equal to 5).

Considering the percentage of clusters that intersects an extreme zone and are classified as extreme, the performance of spatially homogeneous clusters has been superior to the performance of ordinary clusters at all scales.

4. The HR-DASH Protocol

To enable the use of the spatially homogeneous clusters and address the issues of energy saving, we present the concepts and the requirements for the HR-DASH (H*ierarchical* R*outing via* D*ata* A*ggregation with* S*patial* H*omogeneity*) protocol. Our proposal is designed for applications that require data continuously and accept summaries of these data over subregions of the monitoring field.

At the beginning of the network operation, all nodes send their data to the base station. The HR-DASH runs SKATER at the base station, which communicates the nodes about their clusters. After the first data collection, HR-DASH works in rounds as LEACH (Heinzelman et al., 2002). Each round has a set-up phase, when SKATER design the clusters, followed by a steady-state phase. In this phase, nodes collect data continuously but only communicate *extreme changes* to the base station and the head node. There are two reasons to do this. Either the database has to be updated, allowing the user to determine if some important event occurred, such as a high temperature peak. Alternatively, the base station must decide if these changes are large enough so that SKATER should *redesign* the spatially homogeneous clusters. The HR-DASH uses the *statistical properties* of the data to classify a value change either as a relevant one or not.

Changes that are not extreme, but important, are communicated only to the head node to allow for local monitoring. Many such changes detected over a single cluster signal that something is happening and the base station should be reported. Thus, the head node sends the cluster summary to the base station when there is a relevant change in its value. If a node does not send its data to its head, the missing data is replaced by the mean value the cluster head receives from the base station in the set-up phase. Given the cluster internal homogeneity, the cluster mean is a good estimate for the measures. Besides, the missing data replacement allows continuous evaluation of the cluster mean.

The HR-DASH strategy of communicating to the base station only the extreme changes allows for energy saving, because it reduces the number of these costly transmissions. The number of transmissions inside a cluster is also reduced, since nodes send their data to cluster head only if there are relevant changes in their values.

We can configure the HR-DASH protocol to partition the nodes according to a criterion of minimum homogeneity instead of the usual fixed number of clusters. This allows for using fewer clusters where the values of the variable are similar, which saves energy.

5. Concluding Remarks

Within a few years, miniaturized and networked sensors will have the potential to be embedded in several kinds of environments and allow a continuous monitoring (Elson and Estrin, 2004). Geosensor networks will produce a revolution in our understanding of the environment by providing observations at temporal and spatial scales that are not currently possible. Deciding

how these data will be routed to the base station is crucial, since data routing is an important consumer of energy, the most critical resource of the network.

The main contributions of this paper are three fold. First, we have proposed a data-aware clustering procedure that groups the nodes into spatially homogeneous clusters.

Second, we have compared our clustering proposal to the usual clustering procedure of cluster-based protocols. We have shown that spatially homogeneous clusters were able to produce data summaries with a higher statistical quality, improving the posterior statistical analysis. In addition, they get better extreme zones identification.

Finally, we have described the main ideas of the HR-DASH, a protocol to enable the use of our clustering proposal in the data routing. Future work includes the complete specification of the HR-DASH protocol and the evaluation of its energy efficiency.

6. References

Assunção, R. M.; Neves, M. C.; Câmara, G.; Freitas, C. C. Efficient regionalisation techniques for socio-economic geographical units using minimum spanning trees. **International Journal of Geographical Information Science**, v. 20, n.7, p. 797--811, 2006.

Biagioni, E.; Bridges, K. The application of remote sensor technology to assist the recovery of rare and endangered species. **International Journal of High Performance Computing Applications**, v. 16, n.3, 2002.

Chu, D.; Deshpande, A.; Hellerstein, J. M.; Hong, W. Approximate Data Collection in Sensor Networks using Probabilistic Models. In: **International Conference on Data Engineering**. Atlanta, GA, 2006.

Culler, D.; Estrin, D.; Srivastava, M. Overview of Sensor Networks. **IEEE Computer**, v. 37, n.8, p. 41 - 49, 2004.

Elson, J.; Estrin, D., 2004. Sensor networks : a bridge to the physical world. In: Raghavendra, C. S.; Sivalingam, K. M.; Znati, T., eds., **Wireless Sensor Networks**, Kluwer.

Envisense-Secoas, **Self-organizing Collegiate Sensor Networks**, http://envisense.org/secoas.htm.

Goldin, D. Faster In-Network Evaluation of Spatial Aggregation in Sensor Networks. In: **Int'l IEEE Conference On Data Engineering Atlanta**, GA, 2006. p.

Heinzelman, W. B.; Chandrakasan, A.; Balakrishnan, H. An Application-Specific Protocol Architecture for Wireless Microsensor Networks. **IEEE Transactions On Wireless Communications**, v. 1, n.4, p. 660--670, 2002.

Madden, S.; Franklin, M. J.; Hellerstein, J. M.; Hong, W. Tag: a tiny aggregation service for ad-hoc sensor networks. In: **SIGOPS Operating Systems Review**. 2002. p. 31–46.

Mainwaring, A.; Polastre, J.; Szewczyk, R.; Culler, D.; Anderson, J. Wireless Sensor Networks for Habitat Monitoring. In: **ACM International Workshop on Wireless Sensor Networks and Applications**. Atlanta, EUA, 2002. p. 88 – 97.

Martinez, K.; Hart, J. K.; Ong, R. Environmental Sensor Networks. **IEEE Computer**, v. 37, n.8, p. 50 - 56, 2004.

Nittel, S.; Stefanidis, A., 2005. GeoSensor Networks and Virtual GeoReality. In: Nittel, S., Stefanidis, A., ed., **GeoSensors Networks**, CRC Press, p. 296.

Tilak, S.; Abu-Ghazaleh, N. B.; Heinzelman, W. A taxonomy of wireless micro-sensor network models. In: **ACM Workshop on Wireless Security**. 2002. p. 28 -- 36.

Tulone, D.; Madden, S. PAQ: Time Series Forecasting For Approximate Query Answering In Sensor Networks. **Lecture Notes in Computer Science**, n.3868, p. 21--37, 2006.

Warneke, B.; Last, M.; Liebowitz, B.; Pister, K. S. J. Smart Dust: Communicating with a Cubic-Millimeter Computer. **Computer**, v. vol. 34, n.1, p. 44 – 51, 2001.

Xu, N. A Survey of Sensor Network Applications. **IEEE Communications Magazine**, v. 40, n.8, p. 102 – 114, 2002.