DISTRIBUTED ESTIMATION OF STATISTICAL CORRELATION MEASURES FOR SPATIAL INFERENCE IN WSNS

Gustavo Hernández-Peñaloza, César Asensio-Marco, Baltasar Beferull-Lozano

Group of Information and Communication Systems (GSIC)
Instituto de Robótica y Tecnologías de la Información & las Comunicaciones (IRTIC)
Universitat de València, 46980, Paterna (Valencia), SPAIN
Email: {Gustavo.Hernandez, Cesar.Asensio, Baltasar.Beferull}@uv.es

ABSTRACT

This work shows how to obtain distributively important statistical measures such as the semivariogram and the covariogram in a Wireless Sensor Network. These statistics describe the spatial dependence of the sensed area and allow making inferences about unknown field data. In practice, these are complex measures that require global knowledge such as the distance between every pair of nodes, which is not available in a distributed scenario. Then, motivated by the distributed nature of a Wireless Sensor Network and the requirement of making estimations in many real applications, we propose a distributed method to obtain an approximation of these measures, based only on the local samples of the nodes. Our method only requires knowing, at each node, the geographic position of its neighbors. Additionally, we show that introducing random movements of the nodes, the quality of the results can be improved. Simulation results are presented to evaluate the performance of our algorithm.

Index Terms— Statistical tools, Distributed estimations, Wireless Sensor Networks.

1. INTRODUCTION

In Wireless Sensor Networks (WSNs), spatial interpolation techniques such as splines and Kriging [1][2], have attracted a great deal of research work because of their relevance in most of the applications where field reconstruction is required. The goal pursued by these techniques is to acquire global maps [3] of the field behavior, based on the available sensor samples. For example, these techniques are widely used on environmental monitoring [4][2] and spectrum cartography for Cognitive Radios [3].

The aforementioned techniques require some statistical measures [1][6] to exploit the spatial dependence on the field and make inferences about it. The most common statistics employed for this purpose are: the covariance, the correlation, and the semivariance [7]. These tools describe the spatial dependence of the field as a function of the distance. This is done by means of extracting important information of similarity among the samples, at the expense of using a great deal global information such as the distance between every pair of nodes.

In the literature, there exist several methods that consider the problem of spatial interpolation in WSN, employing these tools. For example, in work [4], a method to estimate environmental parameters is proposed, using a centralized covariogram estimation. In [3], a framework to apply spatial interpolation techniques for spectral cartography in Cognitive Radios is presented. However, the entire procedure is performed offline and the statistics involved are assumed to be known. In [12], the process to obtain the semivariogram by means of performing a quadtree protocol is presented at the expense of using global knowledge of the topology.

Our contribution is to obtain, at every node, the spatial similarity of the field through the iterative construction of the semivariogram and the covariogram. In particular, we propose a novel distributed algorithm to capture the spatial dependence of the field, which is based on the diffusion of information within the network. Additionally, we show that the introduction of random movements improves the computation of the statistical measures, increasing the correlation knowledge of the field at every node. However, some lack of information, at certain distances, can arise, reducing the quality of the results as a consequence. In order to alleviate this problem, the nodes can perform a linear regression of the unknown information.

The remainder of this work is structured as follows: the problem is formulated in Section 2. In Section 3, our distributed proposal is explained in detail. The numerical results of this work are summarized in Section 4. Finally, conclusions and future work are given in Section 5.

This work was supported by the Spanish MICINN Grants TEC2010-19545-C04-04 COSIMA, CONSOLIDER-INGENIO 2010 CSD2008-00010 COMONSENS and the European STREP Project HYDROBIONETS Grant no. 287613 within the FP7 Framework Programme, and the GVA fellowship S. Grisolía GRISOLIA/2009/014.

2. PROBLEM FORMULATION

A WSN can be modeled as a graph $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ composed by a set $\mathbf{V}$ of N nodes and a set $\mathbf{E}$ of links. A link $e_{ij} \in \mathbf{E}$ is established between nodes i and j if their euclidean distance, denoted by d_{ij}, is lesser or equal than certain threshold distance R. We define $\mathbf{h} \equiv \{\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_\ell\}$ as a collection of classes of distance. The size of each class of distance is $\frac{d_{\max}}{\ell}$, where $d_{\max}$ denotes the maximum distance between two points in the deployment area. Thus, $\mathbf{h}_1$ contains the distances in $(0, \frac{d_{\max}}{\ell}]$, $\mathbf{h}_2$ contains the distances in $(\frac{d_{\max}}{\ell}, \frac{2d_{\max}}{\ell}]$ and so on. Based on these classes of distance and the data gathered by each node i, denoted by z_i, we propose a distributed method to obtain at every node the semivariogram and the covariogram, which are re-called here for the shake of completeness.

The semivariogram $\hat{\gamma}(\mathbf{h}_s)$: This statistic describes how the data measurements vary with the distance. This is expressed as follows:

$$2\hat{\gamma}(\mathbf{h}_s) \equiv \frac{1}{|\mathcal{N}(\mathbf{h}_s)|} \sum_{\mathcal{N}(\mathbf{h}_s)} (z_i - z_j)^2;\ \forall s = 1, 2 \ldots, \ell \quad (1)$$

where

$$\mathcal{N}(\mathbf{h}_s) \equiv \left\{ (i,j) \mid \frac{(s-1)}{\ell} d_{\max} \leq d_{ij} \leq \frac{s}{\ell} d_{\max};\ \forall i, j \in \mathbf{V} \right\}$$

is the set containing the pairs of nodes i and j such that the distance between them d_{ij} is included in $\mathbf{h}_s$. Fig. 1 shows an example of $\mathcal{N}(\mathbf{h}_1)$ and $\mathcal{N}(\mathbf{h}_2)$ from the point of view of four different nodes.

The Covariogram $\hat{\mathbf{C}}(\mathbf{h}_s)$: This estimator describes how the correlation between the samples decreases with the distance. This is expressed as follows:

$$\hat{\mathbf{C}}(\mathbf{h}_s) = \frac{1}{|\mathcal{N}(\mathbf{h}_s)|} \sum_{\mathcal{N}(\mathbf{h}_s)} (z_i - \mu_i)(z_j - \mu_j)^T \quad (2)$$

where μ_i and μ_j correspond to the estimated mean in the nodes i and j respectively.

Notice that the mean of the field is not necessary to compute the semivariogram, conversely to the covariogram, where the mean of the field must be estimated. For this purpose, we use the geographic gossip algorithm [8]. This algorithm distributively computes the global mean at every node by exchanging their current estimations. Therefore, for the computation of the covariogram, the nodes send both their field measurement and their current estimation of the mean.

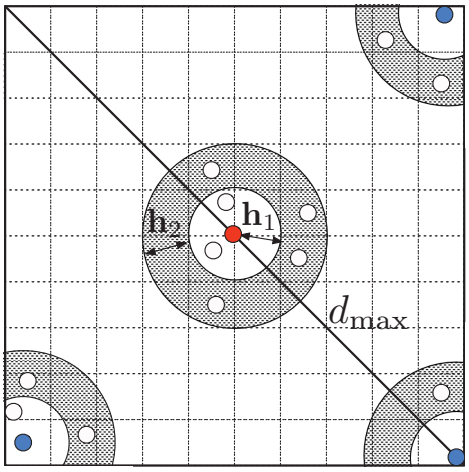


Fig. 1. Example of a unit square area where the nodes have been randomly placed. We show $\mathcal{N}(\mathbf{h}_1)$ and $\mathcal{N}(\mathbf{h}_2)$ for the nodes colored in red and blue. $\mathcal{N}(\mathbf{h}_1)$ corresponds to the nodes inside the circular white areas and $\mathcal{N}(\mathbf{h}_2)$ corresponds to the nodes within the shaded circular areas.

In order to calculate distributively the aforementioned statistic measures, we make the following assumptions:

- Nodes know their own geographical position and the one of their local neighbors.
- We assume that the nodes are randomly and uniformly deployed on the unit square area.
- Additionally, the quality of estimation can be improved if node mobility is introduced. If mobility is allowed, the new positions of the nodes are independently selected, as in [11].

Lemma 1: If the number ℓ of classes of distance satisfies $0 < \ell \leq \sqrt{\frac{\pi N}{\log N}}$, then, with high probability, there is at least one node in each class of distance.

Proof: Assuming a unit square area, it has been shown in [10] that a sub-area of size $\frac{2 \log N}{N}$ ensures with high probability, at least, one node inside it. Since the minimum area $\pi \left(\frac{d_{\max}}{\ell}\right)^2$ corresponds to the first class of distance, the parameter ℓ must be at most $\sqrt{\frac{\pi N}{\log N}}$.

Remark: Note that the nodes located around the corners cover smaller areas than the rest of the nodes, (see Fig. 1). However, these nodes, as opposite to the ones around the center, are able to obtain information of the larger classes of distance. Both problems can be solved by introducing random movement of the nodes and simple linear regressions, as we explain in Section 3.

3. DISTRIBUTED APPROACH

The intuition of our method is to exploit the spatial information in a distributed manner, by allowing the continuous diffusion of messages, similar to [8]. This provides essential information to the nodes for the computation of the statistical measures presented in Section 2. Optionally, we introduce the random movement of the nodes in order to improve the quality of the estimations.

In each iteration of our algorithm, the network performs one of the following two operations: 1) a randomly selected node i sends a message containing its own geographic location and its sensed data to some random coordinates or 2) the node moves to some random point in the area, and samples the field again. We denote by $0 \leq \alpha \leq 1$ the parameter that indicates the probability of movement. **Algorithm** 1 describes how the nodes choose between one of the two previous operations with probability α, and **Algorithm** 2 shows how the forwarding process is performed.

Algorithm 1 Origin node i

Require: $x_i, y_i, z_i, \mu_i, \alpha$
  x_d= rand(0:1), y_d=rand(0:1)
  p = rand(0:1)
  if $p > \alpha$ **then**
    message = $[x_i, y_i, x_d, y_d, z_i, \mu_i]$
    choose neighbor j closest to (x_d, y_d)
    send message to neighbor j
  else
    move to (x_d, y_d)
    $x_i = x_d$; $y_i = y_d$
    z_i = senses a new value
  end if

Algorithm 2 Forwarding node j

Require: $[x_i, y_i, x_d, y_d, z_i, \mu_i], x_j, y_j, z_j, \mu_j$
  d_{ij} = obtain distance to origin node i
  $\mathbf{h}_s$ = class of distance corresponding to d_{ij}
  $[\hat{\gamma}(\mathbf{h}_s), \hat{\mathbf{C}}(\mathbf{h}_s)]$ = refine the estimation
  if exists some neighbor closer to destination (x_d, y_d) **then**
    choose the neighbor l closest to (x_d, y_d)
    send message to neighbor l
  else
    ACK message = $[x_j, y_j, x_i, y_i, z_j, \mu_j]$
    $\mu_j = \frac{\mu_i + \mu_j}{2}$
    send ACK message towards (x_i, y_i)
  end if

The first operation implies that a message is sent from the source node to the one hop neighbor that is closest to the random position (x_d, y_d). This operation is iteratively repeated along the route until the current node has no one hop neighbor with shorter distance to (x_d, y_d) than its own. Every intermediate node is able to update its current statistic value with the sample and distance of the source node. In other words, every node in the route computes the distance to the source node by obtaining the coordinates of the received packet and it assigns the obtained value to the appropriate class of distance. Notice that nodes do not require global knowledge of the network topology because the messages are sent to random locations within the deployment area.

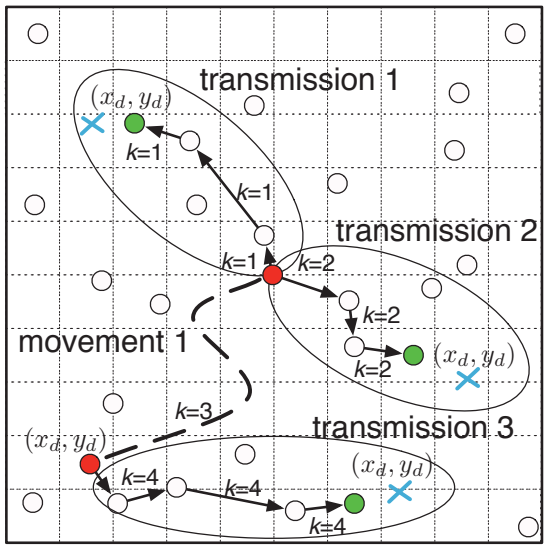


Fig. 2. Example of four iterations of our algorithm centered on the central red node. During these iterations, three multihop transmissions ($k = 1, 2, 4$) and one random movement ($k = 3$) have been performed. The red node is the origin node, some white nodes acts as forwarding nodes and the green ones are the destination nodes of the corresponding transmission.

The second operation implies that a node moves to a new location in the unit square area, it obtains a new measure from the environment, substitutes the previous one and update its mean value. As a result, several values for each class of distance are possibly created in the current node. Additionally, a value in one class of distance is potentially created in the rest of the $N - 1$ nodes. These two operations allow every node in the network to iteratively obtain useful information to construct the semivariogram and the covariogram.

Finally, in order to obtain the covariogram, the average of the samples is required. For this aim, the closest node to destination, use the incoming information to perform an iteration of the geographic gossip algorithm. This allows them to distributively obtain an approximation of the field data mean. A gossip algorithm requires symmetric communication links in order to assure convergence to the data mean. Likewise, the destination nodes use the origin coordinates to send back the message to the source. In the case of introducing random

movements in the network, a node that is waiting for an ACK message is not allowed to move until the message is successfully received.

An example of these operations is represented in Fig. 2. In this example, four iterations are shown from the point of view of node i. In the first and second iterations, the node i sends a message to some random locations (x_d, y_d). By forwarding the message, the nodes in the route (the white nodes in the example shown in Fig. 2) are able to refine the semivariance with the origin node i, at the corresponding class of distance. In the third iteration, node i moves to other new generated coordinates within the deployment area and senses a new value. Finally, in the fourth iteration, the node i sends a new message to a new location allowing the acquisition of new information at another node.

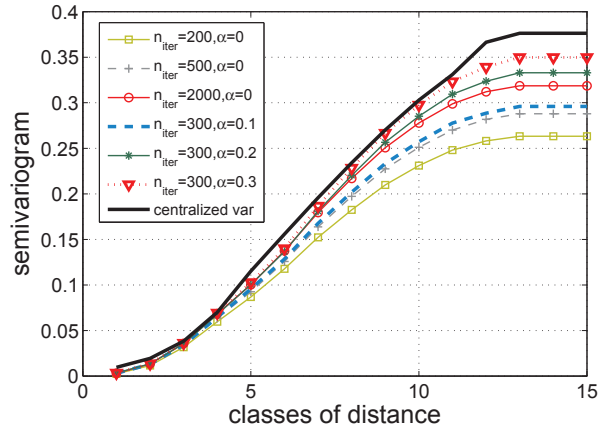
4. NUMERICAL RESULTS

In our simulations, the nodes are randomly deployed over the unit square area with a range of communication of $R = 0.2$ for each of them. The results are obtained by averaging over 100 different random networks of size $N = 100$. In order to evaluate the performance of our distributed algorithm, we create a correlated Gaussian field following the model presented in [13]. Finally, we scale the number of classes of distances ℓ as $\sqrt{\frac{\pi N}{\log N}}$ to ensure that there exist at least one node at each lag of class of distance.

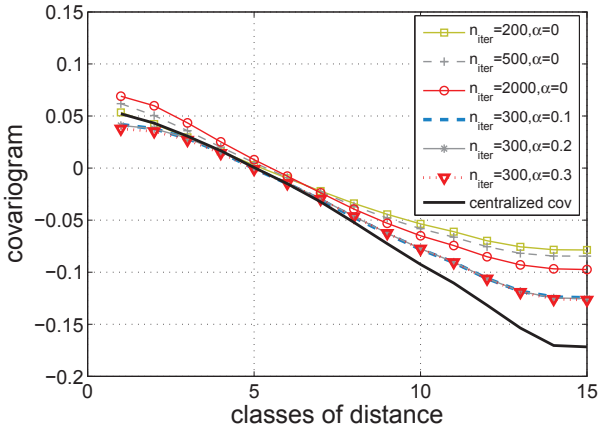
In Fig. 3, we can observe the results obtained for the semivariogram and the covariogram, respectively, as a function of the classes of distance. For the semivariogram Fig. 3(a), when the network nodes are not allowed to move ($\alpha = 0$), the parameter that governs the performance is the number of iterations. The higher the number of iterations is, the better the quality of estimation. Furthermore, we introduce some random movement in the network nodes by slightly varying the parameter α. The results obtained when movements are considered, outperforms the ones given by a static network. This occurs even in the case of performing a small number of iterations, allowing the semivariogram estimation to converge more accurately to the centralized one.

Moreover, in the covariogram case Fig. 3(b), for the static network, improvement is obtained by minimizing the error of the measure with respect to the centralized estimation. Likewise to semivariogram case, the introduction of random movements improves the performance with respect to the centralized covariogram. However, effectiveness provided by the random movements is not as important as the one obtained by the semivariogram estimation.

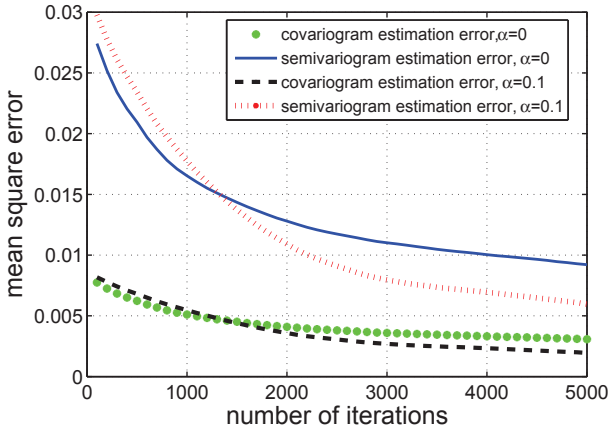
Finally, Fig. 3(c) shows how the mean square error decreases with respect the centralized algorithm as a function of the number of performed iterations. This evolution is shown for different values of the parameter α.



(a)



(b)



(c)

Fig. 3. (a) Semivariogram and (b) covariogram as a function of the classes of distance for different number of iterations. Furthermore, random movements are added for several values of α. All results are compared with the statistic obtained in a centralized manner and the curves of the corresponding mean square error are shown in (c).

5. CONCLUSIONS

In this paper, we have proposed how to face with the problem of semivariogram and covariogram estimation in a distributed manner. We presented an algorithm that allows each node in the network, to exchange messages and execute interpolations to iteratively build the desired tools. We show that the algorithm performance can be improved by adding random movements in a small subset of nodes, according to a predefined factor. These changes on the topology increase the available knowledge of the field, giving as a result that the number of required iterations to obtain a good estimation is reduced. As future work, we will focus on extend the algorithm development to make it adaptive for estimation of the statistics in nonstationary fields.

6. REFERENCES

[1] Martinez, S.; , "Distributed Interpolation Schemes for Field Estimation by Mobile Sensor Networks," *IEEE Transactions on Control Systems Technology*, vol.18, no.2, pp.491-500, March 2010.

[2] Hernández-Peñaloza G. Beferull-Lozano B. "Field Estimation in Wireless Sensor Networks Using Distributed Kriging", accepted in *IEEE International Conference on Communications*, ICC 2012, Ottawa June 2012.

[3] Wellens, M.; Riihijarvi, J.; Gordziel, M.; Mahonen, P.; , " Spatial Statistics of Spectrum Usage: From Measurements to Spectrum Models," *IEEE International Conference on Communications*, ICC 2009, vol., no., pp.1-6, 14-18 June 2009.

[4] Dardari, D.; Conti, A.; Buratti, C.; Verdone, R.; , "Mathematical Evaluation of Environmental Monitoring Estimation Error through Energy-Efficient Wireless Sensor Networks," , *IEEE Transactions on Mobile Computing*, vol.6, no.7, pp.790-802, July 2007.

[5] Schwarz, V.; Matz, G.; , "Distributed reconstruction of time-varying spatial fields based on consensus propagation," *IEEE International Conference on Acoustics Speech and Signal Processing*, ICASSP 2010 , pp.2926-2929, 14-19 March 2010.

[6] Rui Dai; Akyildiz, I.F.; , "A Spatial Correlation Model for Visual Information in Wireless Multimedia Sensor Networks,", *IEEE Transactions on Multimedia*, vol.11, no.6, pp.1148-1159, Oct. 2009.

[7] N.A.C. Cressie .; "Statistics for spatio-temporal data", : *John Wiley & Sons*, 624 pp, New York (2011).

[8] Dimakis, A.D.G.; Sarwate, A.D.; Wainwright, M.J.; , "Geographic Gossip: Efficient Averaging for Sensor Networks," , *IEEE Transactions on Signal Processing*, vol.56, no.3, pp.1205-1216, March 2008.

[9] Zhenghao Zhang; Husheng Li; Changxing Pei; , "Optimum experimental design for estimating spatial variogram in cognitive radio networks," *44th Annual Conference on Information Sciences and Systems*, CISS 2010, vol., no., pp.1-6, 17-19 March 2010.

[10] A. E. Gamal, J. Mammen, B. Prabhakar, and D. Shah, Throughputdelay trade-off in wireless networks, *IEEE Communications Society in Proc. 24th Confeference*, INFOCOM 2004.

[11] M. Grossglauser and David N. C. Tse. 2002. Mobility increases the capacity of ad hoc wireless networks. *IEEE/ACM Transactions on Networking* vol.10, no.4, August 2002.

[12] Muhammad U, Lars K, and Egemen T; Kriging for Localized Spatial Interpolation in Sensor Networks, *In Proceedings of the 20th international conference on Scientific and Statistical Database Management*, SSDBM 2008.

[13] Apoorva Jindal and Konstantinos Psounis; "Modeling spatially correlated data in sensor Networks" *ACM Transactions on Sensor Networks*, 2006, pp.466-499.