A visit design of a data collector robot in a wireless sensor network

Marcia Helena Barbian*
UFRGS/UFMG, Porto Alegre, Brasil - mhbarbian@gmail.com

Renato Martins Assunção
UFMG, Belo Horizonte, Brasil - assuncao@dcc.ufmg.br

Andrea Iabruti Tavares
UFOP, Ouro Preto, Brasil - andrea.iabrudi@gmail.com

Abstract

The purpose of a sensor network is to monitor physical or environmental characteristics in a region such as temperature, pressure or humidity. Due to the great difficulty of replacing batteries of sensor node, power consumption becomes a critical factor in the network. The main energy demand is due to the transmission data. In this context, this project proposes a robot visit only part of sensors and perform data collection, avoiding transmission over long distances. What will contribute considerably to the extension of the life of the entire network. On the other hand, reducing the sample size would increase the variance of prediction. This is a sample design problem in which it is necessary to determine a subset of sensor nodes, that provide estimates with minimal variability. In this paper, we present two algorithms that seek to choose a subsample of sensor nodes that provide minimal variability in prediction process. Furthermore, the performance of the proposed algorithms are compared empirically using Monte Carlo trials.

Keywords: sensor node; subsample; design; prediction.

1. Introduction

The investigation of a phenomenon is usually performed by collecting his observations. There are cases where that amount has to be observed in some region D that is dangerous or difficult to access. In the environmental area, this type of difficulty can be found, for example, in monitoring the temperature in forest regions or in monitoring the level of radioactivity in nuclear disaster spots. This kind of problem is not limited to the environment science and can occur in the military, industrial, traffic engineering, among others. In these situations, the wireless sensor network becomes the natural way to carry out the monitoring of the phenomenon under study.

A network of sensor node consists of autonomous sensors distributed spatially. The main components of this type of mechanism are the sensor, the observer and the phenomenon. The sensor is the element responsible for monitoring the phenomenon of interest Z. The amount Z is an indexed field in the random region D, ie $Z = Z(s) \in D$ where $D \in R^2$. The observer is the one that want to investigate and get answers about the phenomenon Z.

A sensor node is a node in a wireless sensor network which is able to collect information, perform some processing and communicate with other nodes connected to the network. Essentially, sensors are computers extremely basic in terms of interface and components. Its size can vary from a shoe box to a grain of sand. A key advantage of sensors is its cost, which can range from a few dollars to hundreds of dollars depending on the complexity and size.

Due to the great difficulty of replacing batteries, power consumption becomes a critical factor in the network. The energy used for sending a few kilobytes could be used to perform millions of calculations of the processor. In addition, the power consumption increases with distance from the sensor to the receiver node. For this reason, the main energy demand is due to transmission data. Therefore, efficient consumption protocols are needed to extend the life of the system. One way to save energy of the sensor node is to avoid it to transmit data to distant sites. In this context, this project proposes that a robot visit the sensors and collect the data, avoiding the transmission of data over long distances.

However, the time needed for the robot to collect all the information would be very large, as it should go a long way to visit all sensor nodes. An alternative is to make the robot to visit only a portion of the sample (sensor nodes), which, in turn, increases the variance of the prediction process. Thus, the above problem can be solved through a sample design, in which it is necessary to determine the subset of sensor nodes that provide minimal variability for the predictor of Z.
The methodology proposed in this project will contribute significantly to the extension of the life of all of sensor network, excessively increasing the amount of information collected by them, given that the energy that would be spent in data transmission, can be redirected to the information collection.

2. Objectives
Taking the previous descussion into consideration, the main objective is to choose a subsample of sensors to provide minimal variability in the prediction process of non observable values of Z. The most naive approach is enumerate all possible sub-samples and choose the one that provides the lowest variability. However, depending on the amount of sensor nodes, and size of the subsample, this approach becomes unfeasible. For this reason, we highlight the specific objectives of this project:

- the development of a criterion for choosing the " best " design, in the sense discussed above;
- the design of algorithms for selecting sensor nodes that will be part of the subsample;
- evaluating the performance of the algorithms that will be proposed.

3. Literature review
In recent years, many methods for choosing optimal designs monitoring networks have been proposed in the literature. In general, these methodologies consider some criteria that should be optimized, ie some objective function. Given this function, numerical optimization methods are used for the search of the optimal design, ie the one that optimizes the criterion in question.
In this context, studies conducted by Royle and Nychka (1988) proposes the definition of the subsample in terms of the geometry of observational locations. Informally, the sub-sample is formed by minimizing the distance between the observations and the locations where the predictions are necessary.
Some authors have used Monte Carlo algorithms computationally sophisticated, such as simulated annealing, to seeking the best sampling design. For example, Van Groenigen *et al* (1999) suggests that the subsample minimize the variance of kriging, Ruiz-Cardenas *et al* (2010) propose to use the design that maximizes the entropy of the monitored sites and Ruiz Cardenas *et al* (2012) using Monte Carlo evolutionary Markov chains to determine the subsample that maximizes the expected utility.
Finally, we can mention Diggle and Lophaven (2006), who recommend the design that minimizes the prediction variance *a posteriori*. Through this methodology it is possible to proceed, together with the parameter estimation and prediction in the phenomenon observed locations.

4. Methodology
Geostatistics refers to the spatial statistics field in which the data consist of a finite sample values relating to spatially continuous phenomenon. Formally, consider a region $D \subset R^2$, in geostatistics, the sample is formed by observing a random process $Z(s), s \in D$ on a finite set of locations $s_1, s_2, \ldots, s_n$ the region D. In other words, the sample obtained is $\mathbf{Z} = (Z(S_1), Z(s_2), \ldots, Z(s_n))^t$. A widely used model is the Gaussian linear, the simplest definition is given below:

$$Z(s_i) = \mu + \eta(s_i), \tag{1}$$

where for all $s \in D$, $\eta(s)$ has normal distribution with mean $E[\eta(s)] = 0$ e $Var[\eta(s)] = \sigma^2$. The correlation between two locations s e s' is

$$\text{Corr}[Z(s), Z(s')] = \text{Corr}[\eta(s), \eta(s')] = \rho(h),$$

where $h = s - s'$. That is, the correlation between two locations depends only on the distance between them. Under these assumptions, the joint distribution of the vector $\mathbf{Z}$ is multivariate normal with mean

$\mu\mathbf{1}_n$ and variance-covariance matrix $\sigma^2\mathbf{R}$, where $\mathbf{1}_n = (1, 1, \ldots, 1)^t$ with size n and $\mathbf{R}$ entries is given by $\mathbf{R}_{ij} = \rho(s_i - s_j)$, $i, j = 1, 2, \ldots, n$.
Given the vector of observations $\mathbf{Z}$, one of geostatistics objectives is predicting $Z(s_0)$ for any location not observed s_0. In this context, one of the most used methods is the kriging. In this approach, the predictor of $Z(s_0)$ is given by its conditional expectation given $\mathbf{Z}$. Throughout the text will be considered the case in which the model parameters are known, called ordinary kriging. In this context, $(Z(s_0), \mathbf{Z}^t)^t$ follows a multivariate normal distribution with mean vector $\mu\mathbf{1}_{n+1}$ and variance-covariance matrix given by

$$\begin{bmatrix} \sigma^2 & \sigma^2\mathbf{r}^t \\ \sigma^2\mathbf{r} & \sigma^2\mathbf{R} \end{bmatrix},$$

where $\mathbf{r}$ is a vector with elements $r_i = \rho(s_0 - s_i)$, $i = 1, 2, \ldots, n$.
Through some properties of the multivariate normal distribution, it is possible to show that the minimum mean square error predictor of $Z(s_0)$ is given by

$$\hat{Z}(s_0) = \mathrm{E}[Z(s_0)|\mathbf{Z}] = \mu + \mathbf{r}^t\mathbf{R}^{-1}(\mathbf{Z} - \mu\mathbf{1}_n),$$

and its prediction variance is

$$\mathrm{Var}[\hat{Z}(s_0)] = \sigma^2 - \sigma^2\mathbf{r}^t\mathbf{R}^{-1}\mathbf{r}.$$

You can extend the Equation 1 to include a measurement error, obtaining the following Gaussian model

$$Y_i = \mu + \eta(s_i) + \epsilon(s_i),$$

where ϵ_i are independent random variables with normal distribution and mean 0 and variance τ^2 . For details on this and other geostatistical models see Cressie (1993).
The problem addressed by this project can be solved through a sampling design to select the sub-sample to provide the lowest variability for the predictor $\hat{Z}(s_0)$.
The purpose of the monitoring network is to maximize the information collected, satisfying some constraints such as distance, time and cost. At first, this project will consider the distance traveled by the robot is not relevant. Therefore, the objective is to determine the sensor nodes that will be visited by the robot. For example, consider the problem in which 400-sensors are launched by an aircraft over a region. By the arguments already mentioned only 100 sensor nodes will be visited by the robot. We seek to find the optimal sample design in this case.
First, you must define the meaning of great. In this case, an objective function that measures the predictive quality of the model presented in Equation 1 given a set of observations will be set. Possible criteria may be cited the entropy of the monitored sites (Ruiz-Crdenas *et al.*, 2010), prediction variance to a posteriori (Diggle and Lophaven, 2006), or some geometric criterion.
Let G the monitored sensor nodes array and U those not monitored. Therefore, $G \cup U$ is the sample with all sensor nodes. The objective function to be considered is

$$\int E[(\hat{Z}(s) - Z(s))^2|G]ds. \tag{2}$$

For the calculation of Equation 2, it is considered that:

correlation function is exponential parameter ϕ, namely

$$\rho(h) = \exp^{-\phi h}; \tag{3}$$

the distance used is the quadratic

$$s - s' = (x - x')^2 + (y - y')^2$$

where $s = (x, y)$ e $s' = (x', y')$; area of D is the region denoted by A.

Given the above assumptions, the objective function is
$\int E[(\hat{Z}(s) - Z(s))^2|G]ds =$
$= \int \text{Var}[\hat{Z}(s)|G]ds + \int \text{Var}[Z(s)]ds - 2\int \text{Cov}[\hat{Z}(s), Z(s)|G]ds$
$= A\sigma^2 - \sigma^2 \sum_i \sum_j R_{ij}^{-1} \int r_i r_j ds + A\sigma^2 - 2\sigma^2 \sum_i \sum_j R_{ij}^{-1} \int r_j r_i ds$
$= 2A\sigma^2 - 3\sigma \frac{\pi}{2\phi} \sum_i \sum_j \mathbf{R}_{ij}^{-1} \exp^{-\frac{\phi}{2}[(x_i - x_j)^2 + (y_i - y_j)^2]}$.
The next step is to choose the optimal design that minimizes this criterion. The best way to make such a choice would be to list all sub-samples G and choose one that generates the lowest value of the objective function shown in Equation 2. If the amount of sensor nodes is N and the size of the subsamples is n, there are $\binom{N}{n}$ choices of possible subsamples. Therefore, for large values of N, is unlikely to find an algorithm that lists all combinations of fast and efficient manner.
Taking all that into consideration, this project proposes different algorithms to choose the optimal subsample to minimize the value of the criteria presented in Equation 2. These algoritithms are the following:

- random, in which the sub-sample is drawn randomly from sample points available;
- through a grid, wherein a square regular grid is allocated on the observed surface. Within each square of the grid there is a centroid and the chosen observation is one that is closest to the centroid for all grid squares;
- using the intensity of the points where the choice of a sensor node is made through a random selection, wherein the selection probability is proportional to the intensity of that dot, i.e.,

$$p(s_i) = \frac{n_a}{A\lambda(s_i)}, \tag{4}$$

  where n_a is the size of subsample, A is the area and $\lambda(s_i)$ is the intensity at the point s_i.

The quality of the algorithms above will be evaluated using the objective function defined by Equation 2. The algorithm that obtains the lowest value for this function is considered more efficient.
All preliminary results regarding the performance of the proposed methods above are presented in the next section.

5. Preliminary Results

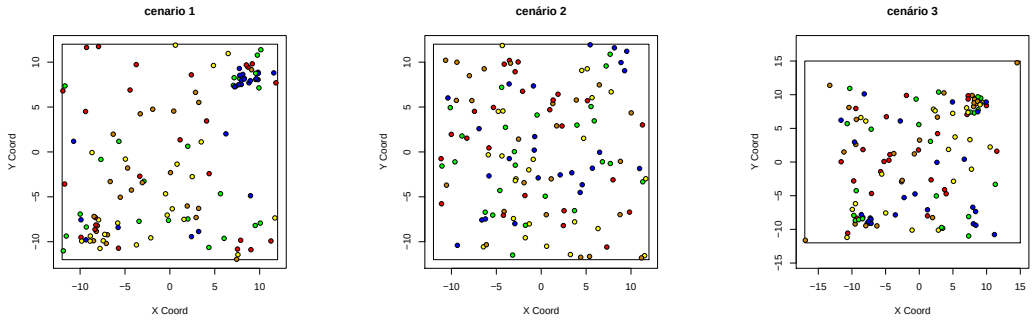
The performance of each of the three algorithms were compared in three different scenarios. In all of them data set of a Gaussian stationary random field with irregularly spaced points were simulated. One hundred replicas we carried out each scenario. The subsamples have size 49, and in all cases, the covariance is the exponential function (Equation 3 with parameter $\phi = 1$). The three scenarios are investigated:

- scenario 1, where the sample size is 122 with two clusters of greater intensity;
- scenario 2, the sample size is 122, intensity is the same for the entire area;
- Scenario 3, the sample size is 124, with two more intense clusters and adding two isolated points.

Figure 1 illustrate some examples of possible scenarios 1, 2 and 3 presented above.
Figure 2 shows the boxplots of the different observed values of the objective function (Equation 2), for each replica Monte Carlo in each of the three scenarios.
From Figure 2, we notice that the grid method performs better in the first two scenarios. In the second scenario, the intensity method of behavior is very close to random. This was expected, due to the fact that in this case, the intensity of the sensors is the same throughout the region D. In the third, the intensity method median below the grid method, but with greater variability.

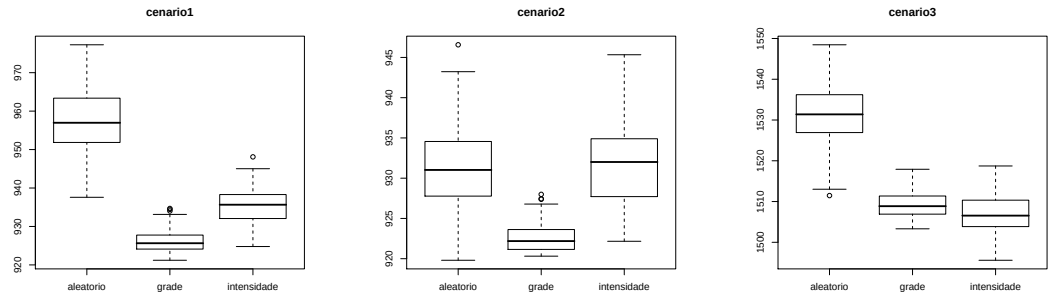
References

Cressie N. (1993) Statistics for Spatial Data. Wiley.



(a) Scenario 1 (b) Scenario 2 (c) Scenario 3

Figure 1: Examples of scenarios investigated.



(a) Scenario 1 (b) Scenario 2 (c) Scenario 3

Figure 2: Box plots of the objective function values in each scenario.

Diggle P. J. & Lophaven S. (2006). Bayesian geostatistical design. Scandinavian Journal of Statistics 33, 55–64.

Royle J. A. & Nychka D. (1988). An algorithm for the construction of spatial coverage designs with implementation in splus. Computers and Geosciences, 24, 479–488.

Ruiz-Cardenas R., Ferreira M.A.R. & Schmidt A. M. (2010). Stochastic search algorithms for optimal design of monitoring networks. Envirometrics, 21, 102–112

Ruiz-Cardenas R., Ferreira M.A.R. & Schmidt A. M. (2012) Evolutionary Markov chain Monte Carlo algorithms for optimal monitoring network designs. Statistical Methodology, 9, 185–194

Van Groenigen J. W., Siderius W. & Stein A. (1999) Constrained optimisation of soil sampling for minimisation of the kriging variance. Geoderma, 87, 239–259.