Design and validation of an optimal network for monitoring the water quality in a hydrological system: A case study

Luis Hernando Hurtado
Gladys Elena Salcedo
María Dolly García

Department of Mathematics. University of Quindio-Colombia
grupopiesa@uniquindio.edu.co

Abstract

The design of environmental monitoring networks is a problem of great practical interest and it has been analyzed from different approaches. In this paper a simple and economical procedure, not only in its design but also in its implementation, is proposed for monitoring the water quality in a hydrological system. The approach consists of an iterative procedure of deletion of sites which is based on the minimization of the cokriging prediction variance. It also considers a stopping rule and a validation method to assess the optimality of the design. An application is made for La Vieja river, Colombia.

Keywords: Monitoring network, spatial correlation, cokriging, comparison of time series

1 Introduction

There exist different methodologies in order to construct optimal monitoring networks and among them there are some specially attractive which are based on the spatial dependence structure of the variables in study represented by the variogram function and use the geostatistical method of kriging.

Most of these procedures consider the addition or deletion of locations to or from an existing network and involves the minimization of an objective function, generally the mean or maximum of the prediction variance. More specifically, they use the fact that the mean-squared prediction error does not depend on the variable value but it depends on its location and the variogram function which can be assumed known or has to be estimated, Carrera et al.(1984), Russo(1984), Warrick and Myers(1987), Spruill and Candela(1990), Ben-Jemaa et al.(1995), Zimmermam(2005). Other proposals as given by Caselton and Zidek(1984) and Ben-Jemaa et al.(1995) use the entropy of the variables in study or the cokriging variance as objective function, respectively.

Some of these before-mentioned works are efficient for spatial prediction, others for estimation of parameters. Diggle and Lophaven(2006) and Zhu and Stein(2006) proposed classical and Bayesian model-based designs integrating both considerations.

Our proposal consists in constructing an optimal monitoring network for spatial prediction from an initial sample, using information of several variables and an iterative process of deletion of sites which is based on the mean-squared prediction error or cokriging variance. It has the advantage that the estimators of the variogram and cross-variogram functions do not change through the process since they can be estimated from the initial sample. We also consider a stopping rule and a validation method that guarantee the efficiency of the design in the sense that the two interpolations obtained from the initial and the optimal network, respectively, are statistically similar.

After applying the methodology to an initial network of 105 sites in La Vieja river (Colombia), it can be concluded that the number of sampling locations might be reduced to 25 in order to study spatially the contamination levels of the river.

In Section 2, the geostatistical method of cokriging is briefly described. In Section 3, the network design approach is presented. The application and the validation of the design are described in Section 4 and finally, some conclusions are given in Section 5.

2 Background theory

Kriging and cokriging are techniques of geostatistical prediction derived from Matheron's theory on regionalized variables and are based basically on the minimization of the mean-squared prediction error.

Consider $k\times 1$ vectors $\mathbf{z}(s_1), \mathbf{z}(s_2), \ldots, \mathbf{z}(s_n)$, where $\mathbf{z}(s_j)$ contains the information of k variables $Z_i,\ i = 1, \ldots, k$, which are sampled at the spatial location $s_j, j = 1, 2, \ldots, n$. The cokriging predictor of the variable Z_1 at a single location s_0, $\hat{Z}_1(s_0)$, is a linear combination of all the $k \times n$ observations $Z_i(s_j),\ i = 1, \ldots, k;\ j = 1, \ldots, n$,

$$\hat{Z}_1(s_0) = \sum_{i=1}^{k}\sum_{j=1}^{n} \lambda_{ij} Z_i(s_j)$$

where the $k \times n$ values λ_{ij} can be obtained by the solution of the following cokriging equations system

$$\begin{aligned} \sum_{i=1}^{k}\sum_{j=1}^{n} \lambda_{ij} C_{ii'}(s_j, s_{j'}) - m_{i'} &= C_{1i'}(s_0, s_{j'}), \quad i' = 1, \ldots, k, \quad j' = 1, \ldots, n \qquad (1) \\ \sum_{j=1}^{n} \lambda_{1j} = 1, \quad \sum_{j=1}^{n} \lambda_{ij} &= 0, \quad i = 2, \ldots, k, \end{aligned}$$

where $C_{ii'}(s_j, s_{j'}) = Cov(Z_i(s_j), Z_{i'}(s_{j'}))$ represents the cross-covariogram of the variables Z_i and $Z_{i'}$ in the sites s_j and $s_{j'}$ for $i, i' = 1, \ldots, k;\ j, j' = 1, \ldots, n$.

The cokriging variance, $\sigma_k^2(s_0)$, measures the uncertainty that is produced when $Z_1(s_0)$ is predicted and is given by the expression

$$\sigma_k^2(s_0) = C_{11}(s_0, s_0) - \sum_{i=1}^{k}\sum_{j=1}^{n} \lambda_{ij} C_{1i}(s_0, s_j) + m_1, \tag{2}$$

where the $\lambda'_{ij}s$ and m_1 are obtained by the solution of system (1).

Introducing the semi-variogram ($i = i'$) and cross-variogram notation

$$\gamma_{ii'}(s_j, s_{j'}) = \frac{1}{2} E[Z_i(s_j) - Z_i(s_{j'})][Z_{i'}(s_j) - Z_{i'}(s_{j'})],$$

where $i, i' = 1, \ldots, k; \quad j, j' = 1, \ldots, n$, the cokriging variance may be also represented as

$$\sigma_k^2(s_0) = \sum_{i=1}^{k}\sum_{j=1}^{n} \lambda_{ij} \gamma_{1i}(s_0, s_j) + m_1. \tag{3}$$

As we can observe in equation (3), the prediction variance does not depend on the variable values but it depends on the semi-variogram function γ_{11} and the cross-variograms $\gamma_{1i},\ i = 2, \ldots, k$, which represent the spatial correlation structure.

When $k = 1$ in equation (1) we have the kriging system and the corresponding kriging prediction variance in equations (2) and (3).

Defining $N(\mathbf{h}) \equiv \{(s_j, s_{j'}) : s_j - s_{j'} = \mathbf{h};\ j, j' = 1, \ldots, n\}$, the semi-variogram (when $i = i'$) and cross-variogram can be expressed as

$$\gamma_{ii'}(\mathbf{h}) = \frac{1}{2} Var[Z_i(s_j + \mathbf{h}) - Z_{i'}(s_j)],$$

and a natural estimator based on the method-of-moments is given by

$$\hat{\gamma}_{ii'}(\mathbf{h}) = \frac{1}{|N(\mathbf{h})|} \sum_{N(\mathbf{h})} \left(Z_i(s_j + \mathbf{h}) - Z_{i'}(s_j)\right)^2,$$

where $|N(\mathbf{h})|$ is the number of distinct pairs of $N(\mathbf{h})$.

As it can be observed $\mathbf{h}$ is a vector with norm and direction. When the cross-variogram function does not depend on the direction of $\mathbf{h}$ we say that the process is isotropic and in this case, we can represent the function as $\gamma_{ii'}(h)$, where $h = \|\mathbf{h}\|$ represents the norm of $\mathbf{h}$. There exist various isotropic semivariogram models in the classical literature of geostatistics, the most known are the spheric, exponential and Gaussian models. These models depend on three parameters called nugget, sill and range. The *nugget (s_0)* represents the micro-scale variations and/or measurement errors. Ideally, the semivariogram increases with the distance from a value equal to the nugget to a constant value that is the *sill (s)*. The *range (r)* is the distance corresponding to the sill; at larger distances the covariance is usually considered to be null. In the application made herein, we used the spheric and gaussian models which are given by the expressions (4) and (5) respectively.

$$\gamma_{ii'}(h) = s_0 + s^2 \left[\frac{3}{2} \left(\frac{h}{r} \right) - \frac{1}{2} \left(\frac{h}{r} \right)^3 \right], \tag{4}$$

$$\gamma_{ii'}(h) = s_0 + s^2 \left(1 - e^{-\left(\frac{h}{r}\right)^2} \right). \tag{5}$$

Further details about regionalized variables and geostatistical theory can be found in Cressie(1993) and Wackernagel(1995).

3 The Procedure

Our procedure to construct a monitoring network for a hydrological system begins by suppose that we have information about a set of variables of interest in N_0 sampling sites which are spatially georeferenced and conforms a set R_0. Furthermore, we assume that this set contains all the information about the system. The purpose is to identify in R_0 a subset R of N sites that contains approximately the same information than R_0 but with $N << N_0$. The subset R thus constructed is called the optimal monitoring network or the *design*.

The approach is based on an iterative algorithm of deletion of sites that is controlled in each stage by the increment of the uncertainty and by the existent relation among the horizon of spatial correlation (h_c), Hurtado et al.(1999), and the distances between the remaining sites. In this procedure, the semi-variogram and cross-variograms are estimated from the initial sample. Hence, they not change through the process.

The procedure of deletion of sites may stop either fixing the number of sites of the network, N, or identifying the minimum number of sites such that the interpolation by kriging will be reliable. The second option is determined when an isolated site of order q is identified. We say that a site s_i is isolated of order q if it has less than q neighbors from a distance smaller than the horizon of spatial correlation (h_c). The choice of q is arbitrary but if we want to reduce considerably the initial network, it has to be small. It can also depend on the resource availability. Nevertheless, since we wish to obtain a design that will be reliable for prediction, we suggest to complement the design with some criterion of validation as given in section 4.3.

The procedure is given by the following iterative algorithm:

1. Eliminate any site of R_0 and calculate its cokriging prediction variance using the $N_0 - 1$ remaining sites.
2. Include again the deleted site.
3. Repeat the steps 1 and 2 for all site of R_0.
4. Identify in R_0 the site with minimum prediction variance.
5. Eliminate the site identified in step 4.

When the size of the network is fixed, using some criterion of specialists, the process may stop after eliminate $N_0 - N$ sites. Otherwise, if we want to optimize the network, the algorithm has to continue the following steps:

6. Calculate the distance between pairs of the remaining sites and verify that anyone of them is an isolated site of order q.

7. Return to 1 considering the fact that the set R_0 has now $N = N_0 - 1$ sites.

   In this case, the procedure finishes after r stages when remainds a set R with N sites such that, in the following step, will appear an isolated site.

4 Application

Our procedure can be applied for any hydrological system except for the validation method which, in this case, is specified for unidimensional process.

4.1 The study zone and the initial network

In order to apply our approach, it is considered a case study to investigate the spatial distribution of polluents and control the contamination levels in La Vieja river, which is located in the west of Colombia, bordering the states: Quindio, Risaralda and Valle del Cauca (see Figure 1). It has a lenght of 102 kilometers approximately and is the main afluent of the Cauca river, the second most important river from Colombia.

Figure 1: Hydrological map of the Colombian West

In its course, the river cross some agricultural, cattle-rising and mining exploration areas and some small industrial and touristic cities establishing the main sources of emission of polluents. Accordingly, the criterion for designing the initial network was allocate the sampling sites looking for the maximum variability among them but without to fall into temptation of sampling only in those places where there were high values of contamination. Under these conditions, 105 georeferenced sites were allocated on the river as it can be seen in Figure 2, and in each site were sampled different variables that are indicators of the contamination. For this application we only consider the variables: Total Suspended Solids, Nitrites and Turbidity Units, labeled in the sequel as Z_1, Z_2 and Z_3, respectively.

Since realizations from a river may be analyzed like an unidimensional series, Figures 3, 4 and 5 show the spatial distribution of data on a cartesian plane, where the horizontal axis (s) represents the distance from the river-head to each site, and the varia-ble values are represented on the vertical axis.

As it can be observed, the variables exhibit some outlier values and a little trend in the direction of water course. This pair of situations are violating the stationarity and normality assumptions. Hence, in Table 1 is specified a corresponding regression model in order to correct the trend, the number of outliers values in each variable and the p-value obtained for the Kolmogorov-Smirnov normality test which is applied after correcting for these outliers.

Table 1. Trend model, number of outliers and p-value of the normality test

Variable	Model	Outliers	p-value
Total Solids	$Z_1 = 61 - 1.61 \times 10^{-3}s + 10^{-8}s^2$	11	> 0.15
Nitrites	$Z_2 = 0.234 + 1.8 \times 10^{-5}s - 10^{-9}s^2$	6	> 0.07
Turbidity	$Z_3 = 7.63 + 9.9 \times 10^{-5}s$	0	> 0.10

4.2 Algorithm to identify the Monitoring Network

Before executing the algorithm described in section 3 in order to identify the optimal monitoring network, the semi-variograms and cross variograms of the variables Z_1, Z_2 and Z_3 have to be previously estimated. Z_1 is the predictor variable. The estimated variograms $\hat{\gamma}_{11}, \hat{\gamma}_{22}, \hat{\gamma}_{33}$ and $\hat{\gamma}_{13}$ are spheric; $\hat{\gamma}_{12}$ and $\hat{\gamma}_{13}$ are Gaussian. Table 2 presents the estimated parameters of each variogram.

Table 2. Estimated parameters of the semi-variograms and cross-variograms

Variogram	Nugget	Sill	Range
γ_{11}	217	400	6500
γ_{22}	0.0001	0.0075	24500
γ_{33}	6	67	24000
γ_{12}	0.2	-2.8	20000
γ_{13}	0	39.5	19200
γ_{23}	0	-0.85	11000

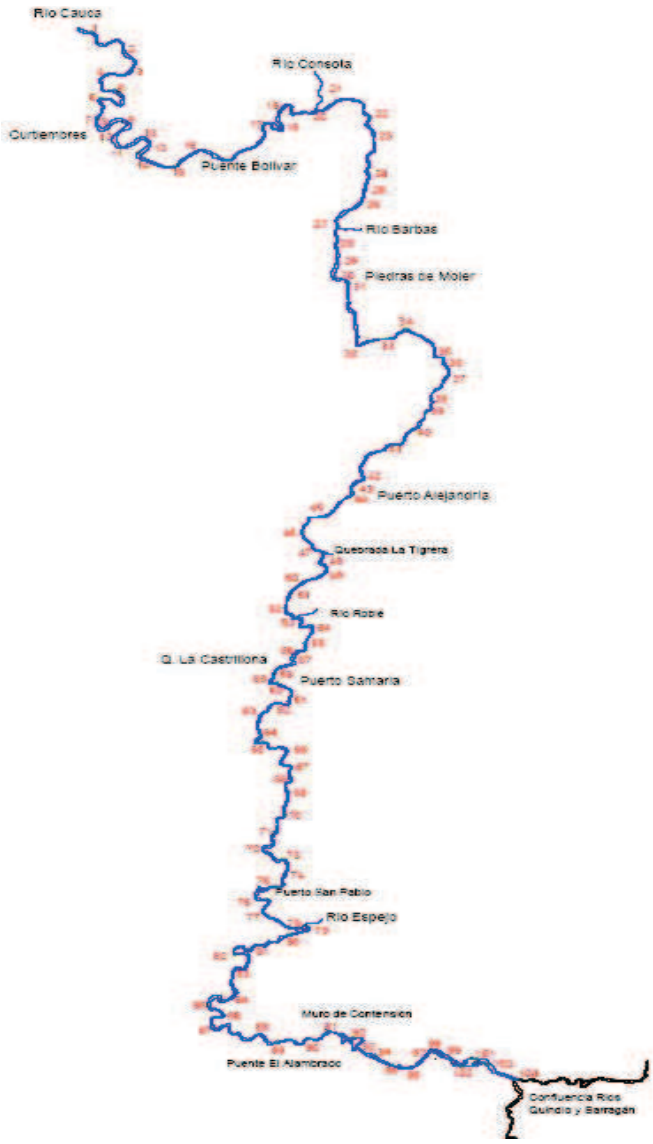

Figure 2: Initial network for La Vieja river

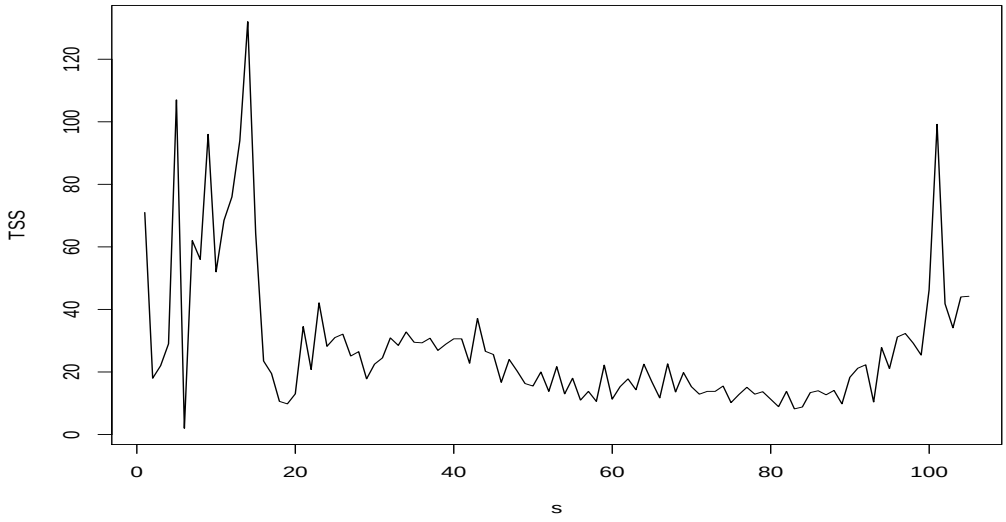

Figure 3: Spatial distribution of Total Suspended Solids

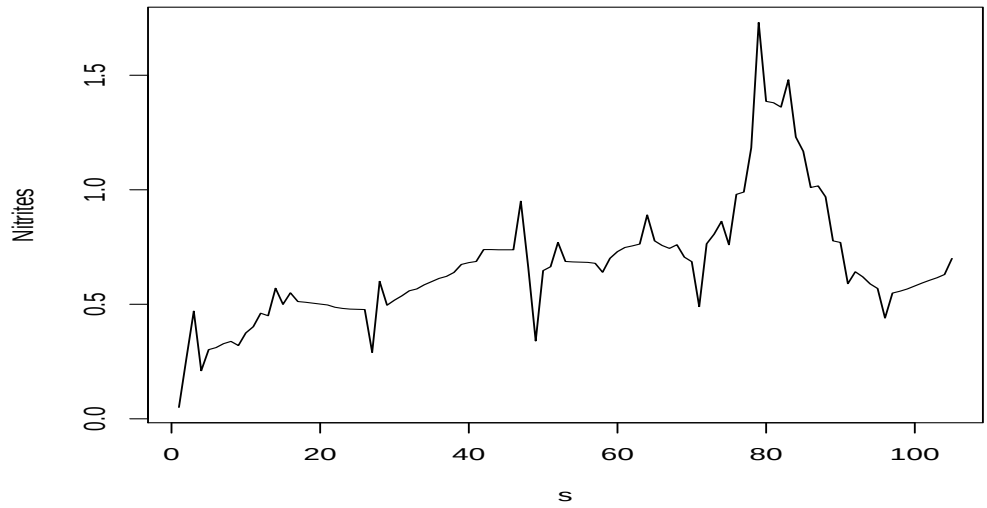

Figure 4: Spatial distribution of Nitrites

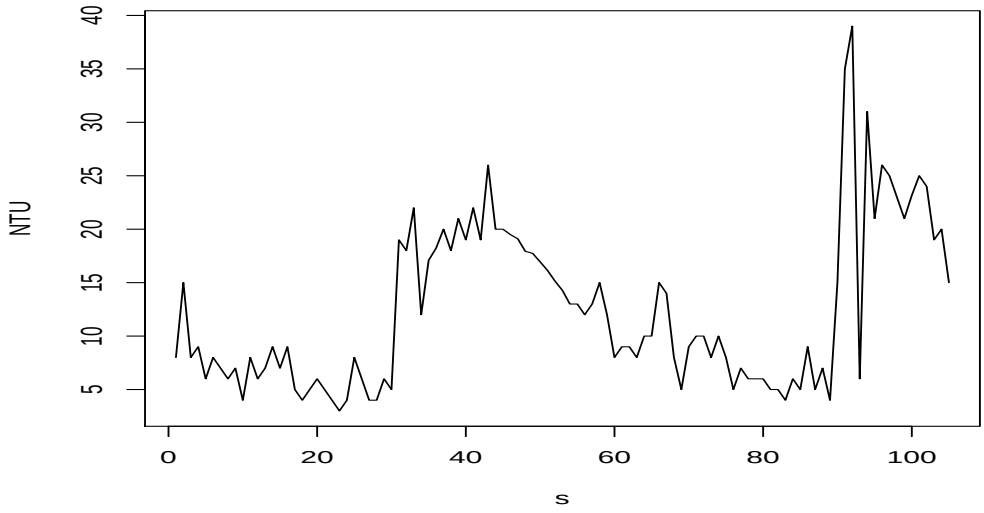

Figure 5: Spatial distribution of Turbidity

After applying the algorithm to the initial sample, with $q = 3$ and $h_c = 10000$ meters, the procedure identified 25 sites distributed as in Figure 6 which corresponds to the optimal monitoring network R_0.

4.3 Validation

The optimal monitoring network thus constructed is in accordance with the expected if we may obtain from R_0 the same prediction that is obtained from R for each variable in study. In others words, if the interpolation made from the 105 initial sites is not significantly different than that made from the 25 optimal sites. Figures 7, 8 and 9 show the two regular interpolations for each variable and, as it was expected, they are very similar except in the places where the variables exhibit extremal values.

There exist differents methodologies to decide if two or more realizations were generated by the same unidimensional process. Thus, in a more formal way, we compare each pair of interpolations by applying some time series comparison test as given for example by Quenouille(1958), Coates and Diggle(1986), Maharaj(2000), Salcedo et al. (2000). Nevertheless, all the procedures above cited suppose that the two series are stationary and free of outliers. Since the outlier values are due to the presence of some pollution focus, we do not consider these observations in the comparison. Hence, we only compare some segments of these series after applying a differentiation of order 1 to transform the series to stationarity ones. For its simplicity, we apply the Quenouille's approach which decides if two independent series, $\{X_t\}$, $\{Y_t\}$, are similar or not by comparing their two corresponding autocorrelation functions. The null hypothesis is

Figure 6: Optimal network for La Vieja river

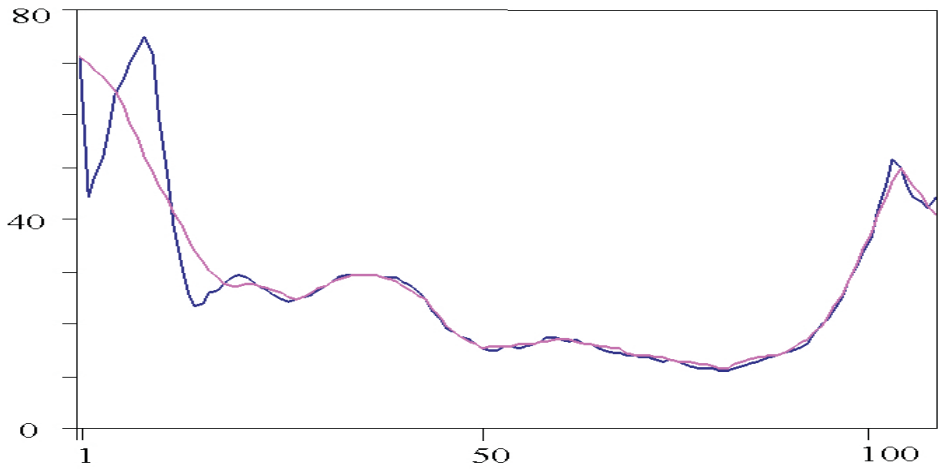

Figure 7: Interpolations obtained from the initial (dark curve) and the optimal network for Total Suspended Solids

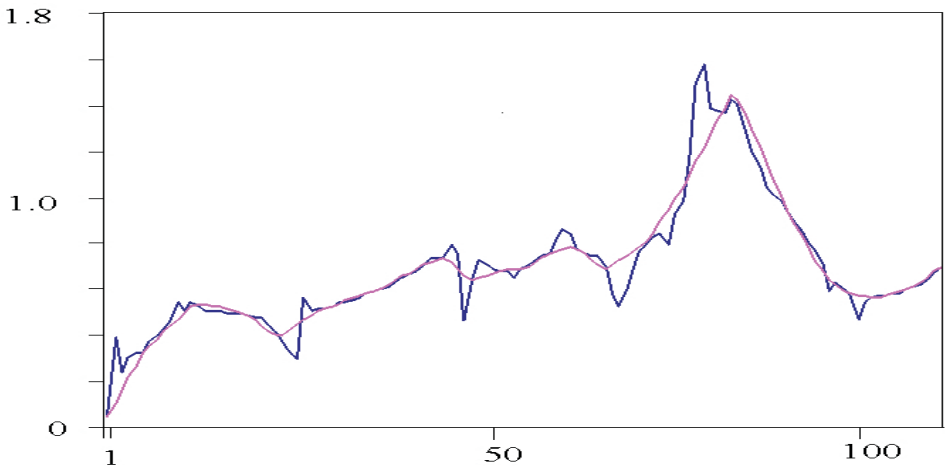

Figure 8: Interpolations obtained from the initial (dark curve) and the optimal network for Nitrites

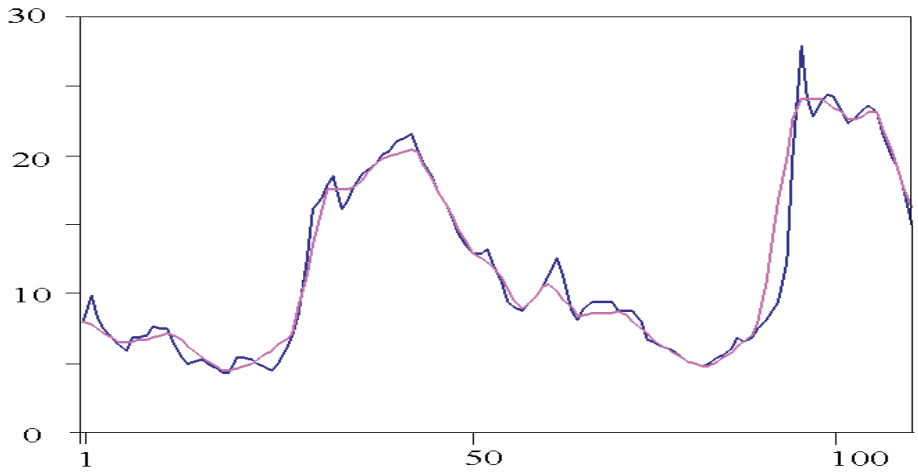

Figure 9: Interpolations obtained from the initial (dark curve) and the optimal network for Turbidity

that $\rho_x(k) = \rho_y(k)$, $\forall k = 0, 1, \ldots, K$, and under H_0, his statistic has asymptotically a χ^2_K distribution. Table 3 shows the p-values obtained when we compare $K = 10$ and $K = 15$ autocorrelations in the segment specified in column 2. Observe that we cannot reject the hypothesis that both series are similar, thus we guarantee the network usefullness. We could also check for similarity in the prediction error variances.

Table 3. p-values of the comparison test

Variable	Segment of Comparison	p-value ($k = 10$)	p-value ($k = 15$)
Total Solids	15-120	0.639	0.878
Nitrits	15-75	0.063	0.229
Turbidity	30-90	0.1963	0.1963

5 Conclusions

We present a simple and economical procedure in order to determine an optimal design for monitoring the water quality in a hydrological system. From an initial network R, the spatial correlation structure of the variables in study is estimated, and using an iterative process of deletion of sites that is stopped when appears an isolated site, we identify a subset R_0 corresponding to the optimal design which is as efficient as R in the sense that it produces the same predictions than R. The application of this procedure to La Vieja river (Colombia) shows that is sufficient 25 sites for monitoring the contamination levels. The spacing between the sites was expected since the network was mainly designed for prediction.

6 Acknowledgements

This work is part of the "*Intercorporation project for the formulation of planning for the management of the hydrological resources of La Vieja river*" which was partially supported by the Autonom Corporations of Quindio, Risaralda and Valle del Cauca. Authors also thank the University of Quindio for its logistic support.

References

[1] Ben-Jemaa, F., Mariño, M.A. and Loaiciga, H.A. (1993) Sampling design for contaminant distribution in lake sediments, *Journal of water Resources , Planning and Management,* **121**(1), 71–79.

[2] Carrera, J., Usunoff, E. and Szidarovsky, F. (1984) A method for optimal observation network design for groundwater management, *Journal of Hydrology,* **73** (1/2), 147–163.

[3] (1984) Caselton, W.F. and Zidek, J.V. Optimal monitoring network designs, *Statistics & Probability Letters,* **2**, 223–227.

[4] Coates, D.S. and Diggle, P.J. (1986) Test for comparing two estimated spectral densities, *Journal of Time Series Analysis*, **7**, 7–20.

[5] Cressie, N.A.C., (1993) *Statistical for Spatial Data,* Jhon Wiley and Sons, New York.

[6] Diggle, P. and Lophaven, S. (2006) Bayesian geostatistical design, *Scandinavian Journal of Statistics,* **33**, 53–64.

[7] Hurtado, L.H., García M.D. (1999) Estimation of the spatial correlation horizon: The case of the variables studied in the Ciénaga Grande of Santa Marta, Colombia. *Bulletin of Investigations. Institute of Marine Investigations of Punta Betin, INVEMAR,* **28**, 158–164 .

[8] Maharaj, E.A. (2000) Clusters of time series, *Journal of Classification*, 297–314.

[9] Quenouille, M. (1958) The comparison of correlations in time-series, *Journal Royal Statist. Soc. Series,* **B. 20**(1), 158–164.

[10] Russo, D. (1984) Design of an optimal sampling network for estimating the variogram,*Soil Science Society of American Journal,* **52**, 708–716.

[11] Salcedo, G.E. and Toloi, C.M. (2000) Tests for comparing time series: Applications to water temperature and salinity measure at different depths, *Brazilian Journal of Statistics*, **61**(215), 51–80.

[12] Spruill, T.B. and Candela, L. (1999) Two approaches to design of monitoring networks, *Ground Water,* **28**, 430–442.

[13] Wackernagel, H. (1995) *Multivariate Geostatistics: An Introduction with applications* Springer Verlag, Berlin.

[14] Warrick, A. and Myers, D. (1987) Optimization of sampling locations for variogram calculations, *Water Resources Research,* **23**, 496–500.

[15] Zhu, Z. and Stein M.L. (2006) Two-step Spatial sampling design for prediction with estimated parameters, *Journal of Agricultural, Biological and Environmental Statistics.* **11**(1), 24-44.

[16] Zimmermam, D.L. (2005) Optimal network design for spatial prediction, covariance parameter estimation and empirical prediction, *Environmetrics.*