Optimal Network Designs for Spatial Prediction

L. Ippoliti*, S. Di Zio, L. Fontanella, R.J. Martin

University G. d'Annunzio, Chieti-Pescara, Italy
ippoliti@unich.it, s.dizio@unich.it, lfontan@unich.it, r.j.martin26@gmail.com
**Corresponding author*

Abstract. *A practical problem in spatial statistics is that of constructing spatial sampling designs for environmental monitoring networks. Among the several purposes for which a monitoring network may be designed for, there is that of interpolation. In this paper, a criterion for spatial designs that emphasize the utility of the network for spatial interpolation of a random field X is discussed.*
Within the Spatial Simulated Annealing (SSA), a stochastic algorithm devised for designing optimal sampling schemes, an R^2 measure criterion is discussed as fitness function. Two different spatial interpolators, namely Kriging and Gaussian Markov Random Fields (GMRFs), are also considered and their potential applications in sampling designs discussed.

Keywords. *Optimal spatial designs; Spatial interpolation; Kriging; Gaussian Markov Random Fields; Simulated Annealing*

---

1 Introduction

Monitoring networks are important to provide information on several environmental aspects such as air pollution, acid rain, water quality, earthquakes etc. For most environmental applications, we also require a prior mapping of the target pollutant agents over the study region so that the construction of an optimal monitoring network becomes a common problem with spatial dependence playing a crucial role.
When a new network is to be constructed, or an existing one augmented or modified, it is important that the monitoring sites are optimally allocated across space to maximize the information available, which can then be used to make reliable and credible inferences about a variable, X, of interest.
When geostatistical data are considered, the monitoring network can be constructed to emphasize the utility of designs for interpolation. Assuming the second-order dependence is known, optimal interpolators (in the sense of minimum mean squared error) are jointly used with design criteria generally expressed as a function of the prediction variance. The specification of the variance matrix $\mathbf{\Sigma}$, and the use of Gaussian Random Fields (GRFs), define an optimal interpolator known in geostatistics as kriging (Cressie, 1993). An alternative approach is to specify a Gaussian Markov Random Field (GMRF), or Gaussian conditional autoregression (Cressie, 1993), which essentially is based on the specification of $\mathbf{\Sigma}^{-1}$.
The purpose of this abstract is twofold. Firstly, it provides a framework for the optimal spatial interpolator in which the two forms of kriging and GMRFs are presented. Secondly, it attempts to provide a new objective function to be used for designing optimal sampling schemes for spatial predictions. Final

remarks on the interpolators, design criteria and computational issues are also provided in the concluding section.

2 Interpolation of Gaussian Fields

In this section we briefly review the essential theory for spatial prediction of GRFs observed in d dimensions. Let $X(\mathbf{s})$ be a GRF, where $\mathbf{s}$ denotes the spatial coordinates of a generic monitoring site to be observed onto a region of interest $S \subset \mathbb{R}^d$. Assume also that L represents a fine lattice with g grid-points overimposed on S, and that the GRF, observed on L, is denoted by $\mathbf{x} = [x(\mathbf{s}_1), \ldots, x(\mathbf{s}_g)]^T$. The assumption of observing X on L can be convenient even when the region is continuous (for example, as the first stage in areal sampling) and it also occurs frequently in practice (for example, in satellite imagery, the pixels can be regarded as representing the interiors of a rectangular lattice).
Suppose that $X(\mathbf{s})$ has mean $\mu(\mathbf{s})$, a parameterised unknown deterministic spatial trend function. We will assume here that $X(\mathbf{s}) - \mu(\mathbf{s})$, is a stationary Gaussian process. The covariance function is defined as $\sigma(\mathbf{h}) = Cov[X(\mathbf{s}), X(\mathbf{s}')]$, where $\mathbf{h} = (\mathbf{s} - \mathbf{s}')$. If $\sigma(\mathbf{h}) = \sigma(\|\mathbf{h}\|)$, where $\|\mathbf{h}\| = \sqrt{\mathbf{h}'\mathbf{h}}$, then the GRF is spatially isotropic and the resulting $(g \times g)$ covariance matrix, $\boldsymbol{\Sigma}$, has elements which are only function of the distances among the sites. In the following we will deal with only isotropic GRFs.
Suppose now that $L = L_1 \cup L_2$, where L_1 contains $u << g$ grid-points for which $X(\mathbf{s}_i)$ is unknown, and L_2 contains the $n = g - u$ sites for which $X(\mathbf{s}_i)$ is known. Assume that it is desired to construct the best linear interpolator for $X(\mathbf{s}_i)$, in a linear least-squares sense, from $\{X(\mathbf{s}_j), j \in L_2\}$. For convenience, assume henceforth that the sites are reordered so that $L_1 = \{n+1, n+2, \ldots, g\}$. Then $\mathbf{X} = \left[\mathbf{X}_o^T \mathbf{X}_m^T\right]$, where $\mathbf{X}_m$ is the $(u \times 1)$ vector of variables to be predicted, and $\mathbf{X}_o$ is the $(n \times 1)$ vector of variables assumed to be observed at the remaining n grid locations. If we further denote with $E[\mathbf{X}_o] = \boldsymbol{\mu}_o$, $E[\mathbf{X}_m] = \boldsymbol{\mu}_m$, $\boldsymbol{\Sigma}_{oo} = Cov(\mathbf{X}_o, \mathbf{X}_o)$, $\boldsymbol{\Sigma}_{om} = \boldsymbol{\Sigma}_{mo}^T = Cov(\mathbf{X}_o, \mathbf{X}_m)$ and $\boldsymbol{\Sigma}_{mm} = Cov(\mathbf{X}_m, \mathbf{X}_m)$, then the required interpolator is obtained as the conditional expectation

$$\tilde{\mathbf{X}}_m = \mathrm{E}[\mathbf{X}_m | \mathbf{X}_o = \mathbf{x}_o] = \boldsymbol{\mu}_m + \boldsymbol{\Sigma}_{mo}\boldsymbol{\Sigma}_{oo}^{-1}(\mathbf{x}_o - \boldsymbol{\mu}_o), \tag{1}$$

$$\boldsymbol{\Xi} = \mathrm{Var}[\mathbf{X}_m | \mathbf{X}_o = \mathbf{x}_o] = \boldsymbol{\Sigma}_{mm} - \boldsymbol{\Sigma}_{mo}\boldsymbol{\Sigma}_{oo}^{-1}\boldsymbol{\Sigma}_{om}. \tag{2}$$

As shown in Fontanella *et al.* (2008), an equivalent formulation of (1) and (2) can be expressed in terms of $\boldsymbol{\Sigma}^{-1}$. Let

$$\boldsymbol{\Phi} = \boldsymbol{\Sigma}^{-1} = \begin{pmatrix} \boldsymbol{\Phi}_{oo} & \boldsymbol{\Phi}_{om} \\ \boldsymbol{\Phi}_{mo} & \boldsymbol{\Phi}_{mm} \end{pmatrix}$$

then

$$\tilde{\mathbf{X}}_m = \mathrm{E}[\mathbf{X}_m | \mathbf{X}_0 = \mathbf{x}_0] = \boldsymbol{\mu}_m + \boldsymbol{\Phi}_{mm}^{-1}\boldsymbol{\Phi}_{m0}(\mathbf{x}_o - \boldsymbol{\mu}_o), \tag{3}$$

$$\boldsymbol{\Xi} = \mathrm{Var}[\mathbf{X}_m | \mathbf{X}_0 = \mathbf{x}_0] = \boldsymbol{\Phi}_{mm}^{-1}. \tag{4}$$

In the Geostatistical approach, equation (1) is known as the kriging predictor, and its use requires the knowledge of $\boldsymbol{\Sigma}$. On the other hand, if $\boldsymbol{\Sigma}^{-1}$ is specified, and u/g is small, equation (3) should be much quicker and more accurate than (1), and avoid the common ill-conditioning problems of $\boldsymbol{\Sigma}$. In general, there is no simple direct specification of $\boldsymbol{\Sigma}^{-1}$ for covariance models. However, the elements of $\boldsymbol{\Sigma}^{-1}$ on a toroidal or infinite lattice are the inverse covariances of the random field X (Fontanella *et al.* 2008; Bhansali and Ippoliti, 2005), so that (3) is then an inverse covariance-based predictor associated with stochastic models for GRFs. For Gaussian processes, as in time series analysis where processes are often approximated by autoregressive models, it may be better to approximate a GRF X by a GMRF Y which has a variance matrix $\boldsymbol{\Psi}$, where we assume that $\boldsymbol{\Psi}^{-1}$ is a (sparse) approximation of $\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Phi}$. Note that in general there is no simple form for the covariances of a GMRF.

3 Optimal spatial sampling schemes and Fitness functions

The notion of optimal design is intuitive and corresponds to the objective of locating n monitoring sites in an optimal fashion over L. Several algorithms have been proposed in literature to address this problem and here we consider the Spatial Simulated Annealing (SSA) (see, for example, van Groenigen and Stein, 1998) which allows designing optimal sampling schemes by minimising a specific fitness function. The choice of an objective function depends on the goal of the sampling and, in general, for prediction purposes, the average or maximum prediction variance of the kriging estimates are considered as reasonable measures of the goodness of a sampling scheme. In this work, a new criterion based on the definition of an *index of linear determinism* is discussed. The index provides a measure of interpolability of the process and, in principle, could be used to characterise the fitness function within the SSA. The index does not depend on the variance of the process, but only depends on the correlation parameter and the sampling design. To build the index we note that from the conditional expectations described in section 2, we may write

$$\mathbf{X}_m = \tilde{\mathbf{X}}_m + \boldsymbol{\zeta}_m$$

where $\boldsymbol{\zeta}_m$ denotes the interpolation error. Since the individual components $\boldsymbol{\zeta}_m$ and $\tilde{\mathbf{X}}_m$ are mutually uncorrelated, it follows that

$$\mathrm{Var}(\mathbf{X}_m) = \mathrm{Var}(\tilde{\mathbf{X}}_m) + \mathrm{Var}(\boldsymbol{\zeta}_m).$$

This last equation decomposes the variability of X, as measured by its variance matrix, as a sum of the variance matrices of $\tilde{\mathbf{X}}_m$ and $\boldsymbol{\zeta}_m$; the former may be thought of as the variability that could be explained from a knowledge of all the neighbours of $\mathbf{X}_m$, and the latter as the unexplained variability due to the interpolation error. Then, let

$$\mathbf{A} \quad = \quad \mathbf{I}_u - \mathrm{Var}(\boldsymbol{\zeta}_m)\mathrm{Var}(\mathbf{X}_m)^{-1} \tag{5}$$

be a normalised measure of association between $\mathbf{X}_m$ and its linear interpolator for which the following inequality, $\mathbf{0} \leq \mathbf{A} < \mathbf{I}_u$, holds. By the matrix inequality, $\mathbf{B} < \mathbf{C}$, we mean that $\mathbf{C} - \mathbf{B}$ is a positive-definite matrix and, specifically, $\mathbf{A} = \mathbf{0}$, if and only if X is a purely random process. Moreover, if $\rho(\mathbf{C})$ denotes a suitable scalar function attached to a matrix $\mathbf{C}$, it follows from the definition of $\mathbf{A}$ that the 'closer' $\rho(\mathbf{A})$ is to $\rho(\mathbf{I}_u)$, the stronger is the association between $\mathbf{X}_m$ and $\tilde{\mathbf{X}}_m$. Hence, $\mathbf{A}$, in this sense, provides information about the linear interpolability of the process.

A multivariate proposal of an index of linear determinism, as a numerical measure of the linear interpolability of a stationary process, should range between zero and one at these extremes and this property is satisfied by the *trace* and *determinant* correlation coefficients defined, for example, as $A_{D_1} = det\,(\mathbf{I} - \mathbf{D})$, which comes directly from equation (5), or as $A_{D_2} = 1 - \{det[\mathrm{Var}(\boldsymbol{\zeta}_m)]/det[\mathrm{Var}(\mathbf{X}_m)]\}$. The latter can be justified by the multiplicative property of determinants, $det(\mathbf{BC}) = det(\mathbf{B})det(\mathbf{C})$, and by noting that $1 - A_{D_2} = det(\mathbf{D})$. An alternative measure of linear interpolability of X may also be constructed by appealing to the additivity property of the trace operator, $tr(\mathbf{B} + \mathbf{C}) = tr(\mathbf{B}) + tr(\mathbf{C})$, and utilising the decomposition of the variance. This leads the following alternative index of linear determinism for a stationary process $A_{T_1} = 1 - \{tr[\mathrm{Var}(\boldsymbol{\zeta}_m)]/tr[\mathrm{Var}(\mathbf{X}_m)]\}$ or, from equation (5), $A_{T_2} = u^{-1}tr\,(\mathbf{I} - \mathbf{D})$.

A comparison of the trace and determinant correlation indices shows that they have a similar form but use different scalar functions of $\mathrm{Var}(\boldsymbol{\zeta}_m)$ and $\mathrm{Var}(\mathbf{X}_m)$. These indices can be used as objective functions to be maximized within the SSA algorithm. Since some of them are easier to use than others, their performance will be tested by a simulation study to assess their computational stability as well as their interpretability.

4 Discussion

This abstract is concerned with the use of Kriging and GMRFs in the context of optimal spatial sampling designs as well as with the proposal of an index of linear determinism as an objective function in the SSA algorithm. Our presentation has briefely focused on the discussion about interpolation theory, and a number of other issues were not addressed. For example, it has been assumed that the dependence structure of the random field X was known. In practice, when a new network is to be constructed, there is only a little prior knowledge of the field such that the use of design criteria expressed as a function of the prediction variance seem to be precluded. In such situations, coverage or space-filling criteria (Nychka et al., 1997; Stevens and Olsen, 2004; Di Zio et al., 2004) that do not involve the covariance function of the process have been proposed as valid alternatives. How to deal with this problem in the context of optimal interpolators will be one of the most important points to be addressed in an extended version of the present abstract.
A further objective of the work aims at exploring the behaviour of the different formulations of the index of linear determinism under different parametrisations of the spatial process. From the computational point of view, if the number of points to be interpolated is small, the interpolation formulae derived from $\mathbf{\Sigma}^{-1}$ require far fewer calculations. Further, if the GMRF is specified using a relatively small set of neighbours, then $\mathbf{\Sigma}^{-1}$ is relatively sparse, so that calculations can be even quicker. All technical details for fitting a GMRF to a stationary GRF will also be discussed in a forthcoming extended work both through a simulation study and with real data examples.

References

[1] Bhansali, R.J., Ippoliti, L. (2005). Inverse Correlations for Multiple Time Series and Gaussian Random Fields and Measures of Their Linear Determinism. *Journal of Mathematics and Statistics*, **1**, 287–299.

[2] Cressie, N. (1993). *Statistics for Spatial Data*, Wiley, New York.

[3] Di Zio, S., Fontanella, L., Ippoliti, L. (2004). Optimal spatial sampling schemes for environmental surveys. *Environmental and Ecological Statistics*, **11**, 397–414.

[4] Fontanella, L., Ippoliti, L., Martin, R. J., Trivisonno, S. (2008). Interpolation of spatial and spatio-temporal Gaussian fields using Gaussian Markov random fields. *Advances in Data Analysis and Classification*, **2**, 63–79.

[5] Nychka, D., Yang, Q., Royle, J.A. (1997). Constructing spatial designs using regression subset selection, in *Statistics for the Environment* Barnett V. and Turkman K.F. (eds.), Vol. 3 Pollutation Assessment and Control. Wiley, New York.

[6] Stevens, Jr. D.L., Olsen, A.R. (2004) Spatially balanced sampling of natural resources. *Journal of the American Statistical Association*, **99**, 262–278.

[7] Van Groenigen, J.W., Stein, A. (1998). Spatial Simulated Annnealing for constrained optimization of spatial sampling schemes. *Journal of Environmental Quality*, **27**, 1078–86.