Geostatistical Analysis under Preferential Sampling

Raquel Menezes[1] and Peter Diggle[2]

[1] Department of Mathematics for Science and Technology, Minho University, Campus of Azurém, 4800-058 Guimarães, Portugal
[2] Department of Maths & Stats, Lancaster University, LA1 4YF Lancaster, UK

Abstract: In geostatistics it is commonly assumed that the selection of the sampling locations does not depend on the values of the spatial variable. One has preferential sampling when this assumption fails (e.g. maximum values search). We first show that the impact of a preferential design on the traditional prediction methods is not negligible. We address this problem by proposing a model-based approach, for stationary Gaussian processes. This new parametric model is founded on a flexible class of log-Gaussian Cox processes. A numerical study is then included to compare the performance of the model proposed and the traditional geostatistical model.

Keywords: geostatistics; marked point process; spatial statistics

1 Introduction and motivation

Suppose that the data for analysis are of the form $(\mathbf{x_i}, y_i) : i = 1, ..., n$, where $\mathbf{x_1}, ..., \mathbf{x_n}$ are locations within an observation region $D \subset I\!R^2$ and $y_1, ..., y_n$ are measurements associated with these locations. The $\{\mathbf{x_i} : i = 1, ..., n\}$ is the *sampling design* and y_i is assumed to be a realization of $Y_i = Y(\mathbf{x_i})$, where $\{Y(\mathbf{x}) : \mathbf{x} \in D\}$ is the *measurement process*. We also assume the existence of an unobserved *field process* $\{S(\mathbf{x}) : \mathbf{x} \in D\}$, usually regarded as our goal of prediction. Often, Y_i can be considered as a noisy version of the underlying random variable $S(\mathbf{x_i})$, the value at location $\mathbf{x_i}$ of process S.

Preferential sampling refers to any situation in which the sampling design process is stochastically dependent of the field process S. Consequently, the corresponding geostatistical model (specified by the joint distribution of the processes involved) must take into account the conditional distribution of the sampling design. For example, some prior scientific knowledge about S, such as the expected local level of contamination in air pollution, may cause the concentration of samples in areas with atypically large values.

1.1 A class of log-Gaussian Cox processes

The sample locations $\mathbf{x}$ are assumed to be realizations of a point process P. Under complete spatial randomness (CSR), the point process modelling is typically based on some homogeneous Poisson process. To propose a model for preferential sampling, we need a class of point processes where the constant intensity λ of the Poisson process is replaced by a spatially varying intensity function, $\lambda(\mathbf{x})$. More precisely, we wish to model aggregated spatial point patterns where the aggregation is due to some stochastic heterogeneity. This leads us to a class of inhomogeneous Poisson processes, P, with stochastic intensity functions, called the Cox processes.
Additionally, we must have in mind our intention to model the dependency of point process P on field process S. Assuming Gaussian data, we shall then consider log-Gaussian Cox processes for P, i.e. Cox processes where the logarithm of the intensity surface is a Gaussian process (see e.g. Moller et al., 1998).
A **geostatistical model for preferential sampling** is a specification of the joint distribution of the field process, the point process and the measurement process of the form $[S, P, Y] = [S][P|S][Y|S]$ where [.] means "the distribution of". So, the proposed model for preferential sampling might be exemplified by:

- $S \sim \text{SGP}(\mu, \sigma^2, \rho(.))$ – S is a stationary Gaussian process with mean μ, standard deviation σ and spatial correlation function $\rho(.)$.
- $P|S \sim \text{Poisson}\left(\exp\{\alpha + \beta S(\mathbf{x})\}\right)$ – α and β are real numbers, and $|\beta|$ identifies the **degree of preferability**.
- $Y(\mathbf{x_i}) = S(\mathbf{x_i}) + Z_i,\ i = 1, ..., n,$ where $Z_i \sim \text{N}(0, \tau^2)$.

The signal of β is an indicator of a positive or negative association between P and S. A null value for β leads to the **classical geostatistical model** and, in this case, one has $[S, P, Y] = [S][P][Y|S]$.

1.2 Prediction ignoring preferential sampling

We now wish to assess how misleading it might be using standard kriging methodology for preferential sampling. Suppose our target of prediction is $S(\mathbf{x_0})$ the value of process S at a generic location $\mathbf{x_0}$, given sample data $(\mathbf{x_i}, y_i), i = 1, 2, ..., n$. The prediction problem may be formalized by invoking the conditional distribution of S given the observed data $\mathbf{y}$. $\text{E}[S(\mathbf{x_0})|\mathbf{y}] = \widehat{S}(\mathbf{x_0})$ specifies the predicted value and $\text{Var}[S(\mathbf{x_0})|\mathbf{y}] = \text{E}[(S(\mathbf{x_0}) - \widehat{S}(\mathbf{x_0}))^2]$ specifies the prediction variance.
Having in mind that kriging provides a BLUE estimator, we proceed with Monte Carlo experiments to examine the behaviour of bias and variance, by analysing the expectation and the variance of the prediction errors, respectively. We compare three sampling designs: CSR, just clustered and

TABLE 1. Impact of preferential sampling on prediction. Monte Carlo approximations for: expectation and variance of prediction errors.

	CSR ($\beta = 0$)	**Clustered**	**Preferential** ($\beta = 2$)
$\widehat{\mathrm{E}}[\mathbf{PE}]$	$(-0.081, 0.059)$	$(-0.082, 0.186)$	$(1.290, 1.578)$
$\widehat{\mathrm{Var}}[\mathbf{PE}]$	$(0.282, 0.363)$	$(0.996, 1.137)$	$(1.150, 1.471)$

preferential. In each case, 500 realizations in all were used, each with $n = 100$. As we are assuming stationarity, we choose just one prediction point, that for which $\mathbf{x_0} = (0.5, 0.5)$. For each replica $j = 1...500$, we derive the corresponding prediction error, $PE_j = \widehat{S}_j(\mathbf{x_0}) - S_j(\mathbf{x_0})$.
The main results are summarized in Table 1. The most evident conclusion is that the bias, represented by $\widehat{\mathrm{E}}[PE]$, considerably increases under preferential sampling. This is because $\widehat{S}(\mathbf{x_0})$ tends to be over-estimated, as we are forcing a higher density of sample locations close to maximum values of the underlying field S. Additionally, it is quite evident, that $\widehat{\mathrm{Var}}[PE]$ increases under preferential or just clustered sampling designs. Bear in mind that this variance represents an empirical prediction variance, which depends on sampling design $\mathbf{x_i}$ and on the estimation of model parameters, and that the latter is affected by a non-random design.
These results illustrate how misleading it would be to adopt classical kriging methodology for preferential sampling. They support the need for an alternative solution, which will be discussed in next Section.

2 A model-based approach

The term *model-based geostatistics* was first used by Diggle et al. (1998) to describe the application of formal statistical models and likelihood-based methods of inference to geostatistical problems. We suggest a solution for the preferability issue following this approach. We aim to apply an explicit parametric stochastic model to preferential sampling, and proceed with likelihood inference to estimate the parameters of the proposed model to allow for spatial prediction. The values of these parameters that maximize the likelihood are referred to as the MLE's.
The marginal likelihood function of the observed data, $L(\theta|\ Y, P)$, is derived from the density $f(\mathbf{y}, \mathbf{x}) = \int f(\mathbf{s}, \mathbf{y}, \mathbf{x}) d\mathbf{s} = \int f(\mathbf{y}) f(\mathbf{x}|\mathbf{s}) f(\mathbf{s}|\mathbf{y}) ds$. In this way, this likelihood can be written as $L(\theta|\ Y, P) = f(\mathbf{y})\ \mathrm{E}_{\mathbf{S}|\mathbf{Y}}\left[f(\mathbf{x}|\mathbf{s})\right]$, and the corresponding log-likelihood function becomes

$$l(\theta|\ Y, P) = \log f(\mathbf{y}) + \log\ \mathrm{E}_{\mathbf{S}|\mathbf{Y}}\left[f(\mathbf{x}|\mathbf{s})\right]. \tag{1}$$

Here, we highlight that the first term is assumed to be the one used under Gaussian assumptions in classical geostatistics (see e.g. Diggle et al., 2003), and the second is the *correction term* obtained for the preferability issue.

TABLE 2. MLE's confidence intervals obtained from a total of 100 independent samples. The true values of model parameters are $\mu = 4$, $\sigma = 1.4$, $\phi = 0.2$ and $\tau = 0.3$. In the case of a non-null degree of preferability, $\beta = 2$.

	Preferential sampling ?			
	No		Yes	
	PS model	Traditional	PS model	Traditional
$\widehat{\mu}$	$(3.880, 4.134)$	$(3.886, 4.141)$	$(3.749, 4.038)$	$(5.090, 5.372)$
$\widehat{\sigma}$	$(1.198, 1.326)$	$(1.211, 1.337)$	$(0.911, 1.046)$	$(0.807, 0.888)$
$\widehat{\phi}$	$(0.165, 0.190)$	$(0.167, 0.192)$	$(0.137, 0.163)$	$(0.112, 0.130)$
$\widehat{\tau}$	$(0.296, 0.322)$	$(0.295, 0.321)$	$(0.296, 0.311)$	$(0.305, 0.318)$
$\widehat{\beta}$	$(-0.015, 0.011)$	—	$(1.752, 1.833)$	—

Numerical study. We now present a numerical study aiming to compare the MLE's assuming the traditional Gaussian model in classical geostatistics with the MLE's obtained under the proposed preferential sampling model. This assessment is implemented for spatial data in both the preferentially and non-preferentially sampled cases. Parameters $\theta = (\mu, \sigma, \phi, \tau, \beta)$ are the target of estimation. The maximization of $l(\theta)$ given in (1) yields the MLE's of the model parameters.
Table 2 summarizes the results of this numerical study. We performed 100 simulations of sample data sets with $\mu = 4$, $\sigma = 1.4$, $\phi = 0.2$, $\tau = 0.3$ and $\beta = 0$, and 100 more simulations for $\beta = 2$. The main conclusions may be summarized as follows. As expected, if spatial data were not preferentially sampled, the traditional and the PS models present very similar MLE results. However, the *correction term* proposed for the preferability issue in (1) proves to be important, when the preferability degree is non-null. Note that the MLE's confidence intervals derived using the PS model do not include the true values of the parameters in the case of σ, ϕ and β, nevertheless they are always closer to those true values than the MLEs of the traditional model.

References

Diggle, P.J., Ribeiro Jr, P. and Christensen, O. (2003). An introduction to model-based geostatistics. In: *J.Moller, ed., 'Spatial Statistics and Computational Methods', (LNS, 173)* , 4-86, Springer, New York.

Diggle, P.J., Tawn,J.A. and Moyeed,R.A. (1998). Model-based geostatistics (with discussion). *Journal of the Royal Statistical Society, Series C*, **47(3)**, 299-350.

Moller,J., Syversveen,A.R. and Waagepetersen,R.P. (1998). Log Gaussian Cox Processes. *Scandinavian Journal Statistics*, **25**, 451-482.