Sensitivity Analysis of Spatial Sampling Designs for Optimal Prediction

N. Farzaneh Kharajoo and M. Jafari Khaledi*

Tarbiat Modares University

Extended Abstract. In spatial statistic, the data analyzed which is correlated and this correlation is due to their locations in the studied region. Such correlation that is related to distance between observations is called spatial correlation. Usually in spatial data analysis, the prediction of the amount of uncertain quantity in arbitrary locations of the area is considered according to attained observations from sampling points. Thus, supposing being certain the sample size, it is necessary to select a sampling design which its observations are attained from the best prediction in mentioned points that is called spatial sampling design for prediction. In this paper, the determination of such design is considered.

For this, suppose that \( Z(\cdot) \) is studied quantity that change on the continual region \( D, D \subseteq \mathbb{R}^d \); and the corresponding data with this quantity is spatially correlated. The spatial data usually observed in \( n \) location \( S = \{s_1, s_2, \ldots, s_n\} \subset D \) as \( z = (z(s_1), z(s_2), \ldots, z(s_n)) \). For modeling such data for analysis, it is usually used a random filed as \( Z(\cdot) = \{Z(s), s \in D\} \). It is often supposed that the random field \( Z(\cdot) \) is Gaussian with the mean and covariance function:

\[
E(Z(s)) = f'(s) \beta,
\]

\[
\text{cov}(Z(s_i), Z(s_j)) = \sigma^2 \rho(s_i - s_j; \zeta) + \tau^2 U\{s_i = s_j\}
\]

where \( U \) denotes the indicator function. By supposing that \( \theta = (\sigma^2, \zeta, \tau^2) \) shows the vector of parameters of the covariance function, we have: \( Z \sim N(X \beta, \Sigma_\theta) \) which \( X \) is a \( n \times p \) matrix from certain amounts as \( X = (f'(s_1), \ldots, f'(s_n))^{tr} \) and covariance matrix is also \( \Sigma_\theta(\text{cov}(Z(s_i), Z(s_j))) = \sigma^2 R_\zeta + \tau^2 I \)

*Corresponding author
which \( R_\zeta = (\rho(s_i - s_j); \zeta) \) is the correlation matrix of observations and \( I \) is the identity matrix. Now the prediction of the uncertain amount of the considered random field in arbitrary point \( s \in D \) is considered.

Hence, the joint distribution \((Z(s), Z)\) is also a multivariate normal distribution. Supposing the being known of the model parameters, i.e. \((\beta, \theta)\), the optimal predictor is as below:

\[
\hat{Z}(s; \beta, \theta) = E(Z(s)|z, \beta, \theta) = x^t\beta + k^t_\theta \Sigma^{-1}_\theta (z - X\beta).
\]

When the vector of regression coefficients is uncertain, the maximum likelihood estimation or the generalized minimum-square estimation \( \beta \) is attained as \( \hat{\beta}_\theta = (X^t \Sigma^{-1}_\theta X)^{-1} X^t \Sigma^{-1}_\theta z \) and will be replaced in the optimal predictor. In this situation, by minimizing the predictor variance \( E(\hat{Z}(s; \theta) - Z(s; \theta))^2 \), the best linear unbiased predictor in arbitrary point \( s \in D \) will be attained which is called the universal kriging. The variance of the kriging predictor is as a criterion of closeness of the predictor amount to its real amount and as below:

\[
M(s; S, \theta) = \sigma^2 + \tau^2 - k^T_\theta \Sigma^{-1}_\theta k_\theta + (x - X^T \Sigma^{-1}_\theta k_\theta)T(X^T \Sigma^{-1}_\theta (x - X^T \Sigma^{-1}_\theta k_\theta)).
\]

Considering this subject, it is possible to use \( M(s; S, \theta) \) for defining the design criterion for prediction. In other words, the design which minimizes its kriging variance average, i.e.

\[
V(S; \theta) = \frac{1}{|D|} \int_D M(s; S, \theta) w(s) ds,
\]

that \( w(s) \) is determined weight in location \( s \) and \( |D| \) denotes the area of \( D \), is called the optimized sampling design for prediction. Then, the optimal design is

\[
S^* = \arg \min_{S \subseteq D} V(S; \theta).
\]

Because the studied region is continual, there are uncounted designs and hence, it is impossible to find the optimal sampling design. So by approximating the region \( D \) by a proper network with \( N \) nodes and suppose that the design \( S \) is \( n \) members’ sub-collection of \( D_N \), the criterion of the kriging variance average of design and optimal sampling design is given by:

\[
V'(S; \theta) = \frac{1}{N} \Sigma_{s \in D} M(s; S, \theta)
\]

\[
S^* = \arg \min_{S \subseteq D} V(S; \theta),
\]
respectively. In general situation, for the proper approximation of the continual region, we shall select the number of nodes as big which is possible to have more designs and determination of optimal design is difficult. In this situation in order to find optimized sampling design, the use of the simulated annealing algorithm could be one appropriate remedy. This algorithm is statistics according physic science and the aim is locating material in a place location with the least energy.

Since the optimal sampling design shall have the least criterion of design, our stimulation studies show that the optimized sampling design for prediction is a design which its points is outspread in region $D$. Also, our studies show that sampling designs for prediction, with little amount of correlation, have cluster status; meanwhile the optimized sampling design is outspread and includes some close points. However, by increasing the intensity of the spatial correlation, there is no more cluster status and the optimized design is completely outspread. Also, we found that when the measurement error is increased, moreover the design is outspread; there are some close points in the design. Meanwhile the correlation intensity and the amount of the nugget effect, has the little influence on optimal sampling designs for prediction; and in this setting, designs are robust to parameter changes. In this article, assuming that the correlation function parameters is known or the preliminary sampling is available, then the optimal sampling designs will be considered. But when these parameters are unknown or the preliminary sampling is impossible to derive, one can use Minimax or Bayes methods in order to determine the optimal design. In the Minimax method, we determine the maximum of design criterion for several amounts of parameters, and then select the design which minimizes this amount. Although, in the Bayes method, the weighted average of design criterion is attained with the weights of the previous distribution; and then we select the design which minimizes this amount as the Bayes spatial sampling design.

**Keywords.** spatial data; prediction; Gaussian random field; sampling design; sensitivity analysis, simulation.
N. Farzaneh Kharajoo
Department of Statistics,
Faculty of Mathematical Sciences,
Tarbiat Modares University,
Tehran, Iran.
e-mail: 

M. Jafari Khaledi
Department of Statistics,
Faculty of Mathematical Sciences,
Tarbiat Modares University,
Tehran, Iran.
e-mail: jafari-m@modares.ac.ir

The full version of the paper, in Persian, appears on pages 1–17.