Methodology of optimal sampling planning based on VoI for soil contamination investigation

Ikumasa Yoshida i)

i) Professor, Department of Urban and Civil Engineering, Tokyo City University, 1-28-1 Tamazutsumi Setagaya-ku, Tokyo 158-8557, Japan

ABSTRACT

This paper proposes a method for optimal sampling planning, i.e., the number and placement of additional sampling for land with arbitrary shape and with existing sampling data, under the assumption of Gaussian random field with respect to a characteristic parameter. A set of optimal locations for additional sampling are evaluated as a solution of optimization problem, in which its objective function is Value of Information (VoI), and the optimization method is Particle Swarm Optimization. Optimal number of sampling is also evaluated by total cost, i.e., sum of observation cost and VoI. The balance of penalties and observation cost determines the optimal number of additional sampling.

Keywords: optimal sampling placement, risk, observation, Kriging, random field

1 INTRODUCTION

Many researchers have studied placement algorithms that give a planning of observation in cost effective way, while maximizing the benefits. The optimal observation placement problem contains two aspects, minimization of the relevant uncertainties (maximization of the accuracy) and minimization of total costs. Traditional measures of uncertainty, such as covariance matrix or information entropy, however, do not depict the significance of uncertainty if the consequence due to the uncertainty is not considered. Raiffa & Schlaifer (1961) describe intensively the theory of Value of Information (VoI) in decision making under uncertainty. VoI can be interpreted to be expectancy of risk reduction or benefit obtained by the information. Nojima & Sugito (1999) propose a Bayes decision procedure model with VoI concept optimizing the process of post-earthquake emergency response in highly uncertain conditions to prevent secondary damage by emergency shut-off of lifeline services. Straub (2013) describes intensively concept and application of VoI in maintenance problem of infrastructures. Pozzi & Kiureghian (2011) discuss the application of VoI-based method to structural health monitoring.

Soil contamination is one of the issues we have to cope with in modern society. We judge the need of contamination remediation measures after soil sampling and its test at several sites. The Guidelines by Ministry of the Environment, Japan indicates detailed investigation scheme for the identification of contamination, in which a basic placement of sampling (sampling grid) is introduced. The sampling placement is shown for the square land without any existing sampling information. We sometimes encounter the case where the land has complicated shape and prior information with respect to contamination is given. This paper proposes a method to obtain optimal sampling planning, i.e., the number and placement of new additional sampling for land with arbitrary shape and with existing sampling information, under the assumption of Gaussian random field with respect to a characteristic parameter. The proposed method is applied to optimal sampling planning in one and two dimensional problems.

2 VALUE OF INFORMATION FOR OPTIMAL SAMPLING PLACEMENT

2.1 Risk of decision making

It is assumed that observation is performed to obtain useful information to make decision by comparing estimator \( x \) with threshold limit value \( x_0 \), e.g., to judge contaminated soil or ordinary soil by comparing poisonous material concentration \( x \) and its threshold limit value \( x_0 \). Statistical test has two kinds of error. A type I error (or error of the first kind) is the incorrect rejection of a true null hypothesis. A type II error (or error of the second kind) is the failure to reject a false null hypothesis. Referring to these error types, we define two types of false decision making.

http://doi.org/10.3208/jgssp.JPN-036 2451
i) Decision error type 1
Judge $x < x_0$ when true $x > x_0$ (e.g., to judge that contamination countermeasure is not necessary when it is necessary actually).

ii) Decision error type 2
Judge $x > x_0$ when true $x < x_0$ (e.g., to judge that contamination countermeasure is necessary when it is not necessary actually).

The probabilities of decision error type 1, 2 are denoted as $P_1$, $P_2$ (e.g., $1 - P_1$). The risk of the decision error can be calculated with penalties per unit area $C_1$, $C_2$ for the decision errors and the probabilities. Naturally we should make decision to take lower risk.

$$J = \sum_i L_i = \sum_i \min(C_iP_{i1}, C_2(1 - P_{i1}))$$  \hspace{1cm} (1)

Suffix $i$ indicates a region for estimation of risk. Total risk is calculated by summing up the risk over the area for the evaluation.

Let’s have an example that we have estimator $x=3$ when threshold limit value $x_0=3$. It is assumed that the estimator involves uncertainty and its mean is 3. It is also assumed that penalties of the error type 1, 2 are 10, 2 respectively. These assumed values are only for the illustration, and do not have any actual meaning. If the estimator is judged to be less than the threshold value, the probability of error is 0.5, and its risk is 5. If the estimator is judged to be larger than the threshold value, the probability of error is also 0.5, and its risk is 1. The former and latter are called as risk 1 and 2 respectively. Since the smaller risk should be taken naturally, we should take risk 2. Fig. 1 shows the risk we should take for estimator of which mean is 0 to 4. It is assumed that the estimator is Gaussian and its standard deviation is 0.4.

When the mean of estimator is 3, the risk 1 and 2 are plotted at 1 and 5 respectively. When the mean becomes small, risk 1 also becomes small, on the other hand risk 2 becomes large. The point $x_c$ that risk 1 is equivalent to risk 2 indicates a threshold value for the judgement under uncertainty. We should judge that the estimator is larger than threshold limit value $x_0$ when the mean of estimator is larger than the threshold value for the judgement $x_c$. The difference between the $x_c$ and $x_0$ expresses safety margin. The threshold for judgement $x_c$ is determined by uncertainty of the estimator and the ratio of penalty 1 and 2, $C_1$, $C_2$.

### 2.2 Quantification of VoI in a Gaussian random field

In general, it is difficult to compute VoI so that MC approach is proposed (Liu et al. 2012; Pozzi and Kluregian 2012; Straub 2013). VoI can be, however, computed easily in updating of Gaussian random field, i.e., Kriging, which is a probabilistic interpolation method (e.g. Christakos 1992; Cressie 1991). It is assumed that observation data at new locations are obtained at each observation step.

$$Z^k = \left\{ z_1^T, z_2^T, \ldots, z^k_T \right\}$$  \hspace{1cm} (2)

where $z_i^T$, $Z^k$ represent observation data at step $k$ and up to step $k$. Mean vectors at three types of places are obtained as follows,

$$\begin{bmatrix} \bar{x}_1^k \\ \bar{x}_2^k \\ \bar{x}_3^k \end{bmatrix} = \begin{bmatrix} \bar{x}_1^0 \\ \bar{x}_2^0 \\ \bar{x}_3^0 \end{bmatrix} + \begin{bmatrix} M_{11}^0 \\ M_{12}^0 + R_1^k \\ M_{13}^0 \end{bmatrix}^{-1} \begin{bmatrix} \bar{z}^k - \bar{x}_1^0 \end{bmatrix}$$  \hspace{1cm} (3)

where $\bar{x}_i^k$ represents a mean vector at places where the observation $Z^k$ is given; $\bar{x}_j^k$ is a mean vector at places where new observation $z_i^k$ will be given; $\bar{x}_j^0$ presents a mean vector at area where decision error risk is evaluated; $R$ is covariance matrix of observation error. The covariance matrices of $x_i$, $x_j$ are given as:

$$M_{ij}^k = M_{ij}^0 - M_{ij}^0 \left( M_{11}^0 + R_1^k \right)^{-1} M_{1j}^0$$  \hspace{1cm} (4)

It is noted that locations of $x_i^n$ are those of observation points at $k+1$ step, and the observation data is not obtained yet. As mentioned above penalty is imposed on false decision-making. The risk can be evaluated from the product of probability of false decision making and the penalty.

$$L(\bar{x}_i^k, \sigma_{ij}^k) = \min(C_iP_{i1}, C_2(1 - P_{i1}))$$  \hspace{1cm} (5)

where, $P_{i1} = \Phi(\beta_i)$, $\beta_i = \frac{\bar{x}_i^k - x_0}{\sigma_{ij}^k}$.
Φ is the standard Normal (Gaussian) cumulative distribution function; \( \sigma^i_{3,} \) is standard deviation of \( k^i_{x,3} \) which can be obtained from diagonal component of covariance matrix \( M_{33}^k \) shown in Eq.(4). The total risk at the decision making area is given by:

\[
\sum_{k} \frac{k_{i} k_{j}}{\sigma^i_{3,} \sigma^j_{3,}} (6)
\]

The decision error risk is reduced by the new information \( z^{k+1} \). After we obtained observation vector \( z^{k+1} \), the mean and covariance matrix of \( x^i_3 \) is updated as:

\[
x^i_{3}^{k+1} = \tilde{x}^i_3 + M^k_{33} M^{k+1}_{33}^{-1} (x^{k+1} - \tilde{x}^i_3) \quad (7)
\]

\[
M_{33}^{k+1} = M_{33}^k - M_{33}^k M_{22}^k (M_{22}^k + R_{22}^{k+1})^{-1} M_{23}^k \quad (8)
\]

Naturally value of the new information \( z^{k+1} \) is not given yet. Therefore \( x^i_3 \) instead of \( z^{k+1} \) is used in Eq.(7).

The expectancy of risk reduction is defined as Vol.

\[
\text{Vol} = E[J^{k+1} - J^k] = E[J^{k+1}] - J^k \quad (9)
\]

The expectancy of risk considering observation data in next step \( z^{k+1} \) is

\[
E[J^{k+1}] = \sum_i L(x^i_3, \sigma^i_{3,}) p(x^i_2) dx^i_2 \quad (10)
\]

Integration with respect to \( x^i_2 \) is required, but it cannot be performed analytically. When dimension of \( x^i_2 \) is high, numerical integration is not practical to implement. Thanks to reproductive property of Gaussian, the numerical integration can be always reduced to one-dimensional numerical integration. Consequently Vol can be calculated easily even if the dimension of \( z^{k+1} \) (the number of additional observation points) is large, e.g., more than 10.

2.3 Optimization of Vol with respect to location of new observation

When the dimension of vector \( z^{k+1} \) is low, it is not difficult to optimize the location of new observation. You can determine the optimal location by evaluating Vol at every possible combination of locations. It is, however, difficult to evaluate them due to “curse of dimensionality” when the dimension of the vector \( z^{k+1} \) is high. In this paper PSO (Particle Swarm Optimization) is introduced to optimize a set of location.
of new observation with respect to VoI. PSO is one of
global optimization methods, which was proposed by
Kennedy et al. (1995). It is said that PSO is a simple
method with a few parameters that users must
determine but efficient for optimization with respect to
real number variables.

3 ONE DIMENSIONAL OPTIMAL
PLACEMENT

It is assumed that there are four existing sampling
data as shown in Fig.2. The values shown in the figure
indicate contamination level. Mean and standard
deviation of the random field are 2.0, 1.0.
Autocorrelation distance is 5.0m in all directions. Area
for evaluation of VoI is shown as "construction" with
the blue line in Fig.2. The threshold \( x_0 \) for the
decision-making of soil contamination is 3.0. Penalty
\( C_1 \) and \( C_2 \) are 10 and 2. Standard deviation of
observation error is 0.1. Based on the existing data,
contamination level and its standard deviation along the
construction line are estimated by using Kriging as shown
in Fig.3(1). The distribution of false decision
risk based on the existing data is also shown as "risk
(prior)" in Fig.3(2). It corresponds to the second term in
right side in Eq.(9).

VoI is evaluated for new observation point at 10 m.
The distribution of posterior standard deviation is shown as "St.Dev.(post)". It is naturally indicated that
the standard deviation around the new observation is
reduced. The risk is evaluated reflecting not only the
standard deviation but also the observed value. It
corresponds to the first term in right side in Eq.(9).
The risk is reduced around the new observation point.
The difference between the prior and posterior risk is VoI
which is indicated by the area colored in the figure.

Optimal placements of additional observation points
are determined such that the VoI is minimized with
respect to the position of the additional sampling points.
PSO (Particle Swarm Optimization) is used for the
minimization. The optimal placement is shown in Fig.4
when four additional sampling is performed. In the
same way, optimal placements are determined for 1 to 7
additional sampling points. Total cost which is sum of
observation cost and VoI can determine optimal
number of additional points. Fig.5 shows the total cost
when observation cost is assumed to be 5 or 3. The
optimal number is 2 for observation cost 5, 6 for
observation cost 3.

4 TWO DIMENSIONAL OPTIMAL
PLACEMENT

An example for two dimensional observation planning
is shown. The area for investigation is a square of 32m X
32m. Mean and standard deviation of the random field
are 1.0, 0.5. Autocorrelation distance is 15m in all
directions. Standard deviation of observation error is
0.1. The threshold \( x_0 \) for the decision-making of soil
contamination is 2. Penalty \( C_1 \) and \( C_2 \) are 10 and 2.

Optimal placements for 3, 5 and 6 samplings are
shown in Fig.6. Guideline by Ministry of the
Environment Japan shows the placement that is rotated
with 45 degree from Fig.3(2). When there is no existing
sampling information, the shape of placement is
geometric, which is imaginable by feeling. It is,
however, difficult to tell the optimal placement by
feeling when there are existing sampling data. Fig.7
shows optimal additional sampling placement when
there are three existing sampling information. Three
cases of observed values are considered at the same
locations of existing sampling. Depending on the
observed values, the optimal placements are determined.

---

![Fig.4. Optimal locations for four additional sampling](image1)

![Fig.5. Optimal number of sampling](image2)
Area around the existing sampling of which observed value is close to 2 are put weight because the threshold value is 2.

5 CONCLUSIONS

This paper proposes an efficient method to obtain optimal sampling (observation, boring) placement based on Value of Information (VoI) in a Gaussian random field. VoI contains cost attributable to the uncertainty to assess the usefulness of observation information considering the consequence due to the uncertainty. The proposed method is applied to additional sampling placement in one and two dimensional problems. Optimal number of sampling is also evaluated from total cost, i.e., sum of observation cost and VoI. The balance of penalties and observation cost determines the optimal number though the penalties are difficult to determine rationally. One of the difficulties in practical applications of VoI lies in the determination of parameters like penalties. This will be future topics to be discussed.

The guideline by Ministry of the Environment Japan indicates sampling plan for soil contamination in a square space without existing sampling. The proposed method, which is consistent with the guideline, can determine the optimal sampling planning for any shape and with any existing sampling.

REFERENCES

1) Christakos, G. (1992): Random Field Models in Earth Sciences, Academic Press Inc.
2) Cressie, N. (1991): Statistics for Spatial Data, John Wiley & Sons.
3) Liu, X., Lee, J., Kitanidis, P., Parker, J. & Kim, U. (2012): Value of Information as a Context-Specific Measure of Uncertainty in Groundwater Remediation, Water Resources Management, 26(6), 1513-1535.
4) Kennedy, J. & Eberhart, R. (1995): Particle swarm optimization, Proc. of IEEE Int. Conf. on Neural Networks, Vol.4, 1942–1948.
5) Nojima, N. & Sugito, M. (1999): Bayes Decision Procedure Model for Post-Earthquake Emergency Response, Optimiz-ing Post-Earthquake Lifeline System Reliability, Proc. of the 5th U.S. Conference on Lifeline Earthquake Engineering, 217-226.
6) Pozzi M. & Der Kiureghian A. (2012): Assessing the Value of Alternative Bridge Health Monitoring Systems, 6th International Conference on Bridge Maintenance, Safety and Management, IABMAS: CRC Press.
7) Raiffa, H. & Schlaifer, R. (1961): Applied statistical decision theory. Boston: Clinton Press, Inc.
8) Straub, D. (2013): Value of Information Analysis with Structural Reliability Methods, Structural Safety, special issue in the honor of Prof. Wilson Tang
9) Sun, N-Z. (1994): Inverse problems in groundwater modelling, Kluwer Academic Publishers