12th International Conference on Applications of Statistics and Probability in Civil Engineering, ICASP12
Vancouver, Canada, July 12-15, 2015

Parameter Study on Optimal Sampling Planning Based on Value of Information

Ikumasa Yoshida
Professor, Dept. of Urban and Civil Engineering, Tokyo City Univ., Tokyo, Japan

ABSTRACT: A Method for determining optimal observation placement in Gaussian random field is formulated based on Value of Information (VoI). The proposed method can reflect not only uncertainty but also the consequence of false decision-making. Optimal number of observation is also evaluated based on total cost which is sum of observation cost and VoI. The usefulness of the method is demonstrated through two kinds of numerical examples which are optimal placement with or without existing observation. Parameter study is performed with respect to autocorrelation distance assumed in the Gaussian field.

1. INTRODUCTION

The optimal observation placement problem contains two aspects, minimization of the relevant uncertainties (maximization of the accuracy) and minimization of total costs. Several measures of uncertainty, such as covariance matrix or information entropy have been used in optimal observation placement problems (Sun 1994). The various norms of the parameter or prediction covariance matrix may be used to express overall uncertainty. For example, Honjo & Kudo (1999) use information entropy to study observation scheme for ground deformation prediction, while Hoshiya & Yoshida (1998) use geometric mean of reduction ratio of standard deviation of model or response parameters of a slope.

New observation information reduces the variance of parameters to be identified, however, quantification of reduction in variance is not enough to answer the question whether the new observation should be performed or not. To answer the question, we need to estimate the worth of the information content in data, i.e., the value of information (VoI) (Raiffa & Schlaifer 1961). VoI can be interpreted to be expectancy of cost reduction or benefit obtained by the information. Nojima & Sugito (1999) propose a Bayes decision procedure model with VoI concept to optimize the process of post-earthquake emergency response in highly uncertain conditions to prevent secondary damage by emergency shut-off of lifeline services. Straub & Faber (2005), Straub (2013) intensively discuss concept and application of VoI in maintenance problem of infrastructures. Pozzi & Kiureghian (20121) discuss the application of VoI-based method to structural health monitoring. Wu et al. (2013) propse decision-making framework for earthquake early warning with VoI concept.

A method, which is a combination of Kriging, Particle Swarm Optimization (PSO) and VoI, is proposed for optimal sampling planning for soil investigation. Kriging is a probabilistic interpolation method in Gaussian field. It is widely known as geostatistics and currently used in many fields. PSO is one of a population based stochastic global optimization method. Kriging is used to estimate spatial distribution of parameter for decision-making, while PSO is used to minimize the VoI with respect to placement of observation locations.

Soil contamination is one of the issues we have to cope with in modern society. Measures such as contamination remediation should be implemented at land which does not meet the designation standard based on the results of a soil

contamination investigation. In this paper, an example of optimal sampling placement in terms of VoI is shown for making more rational decision about the contamination remediation. Sensitivity of auto-correlation distance in Gaussian field is discussed through numerical examples.

2. FORMULATION OF VOI BASED PLANNING

2.1. Gaussian Random Field

Prior information as to random variable vector $\boldsymbol{x}$ is given as,

$$\boldsymbol{x} = \bar{\boldsymbol{x}} + \boldsymbol{w} \tag{1}$$

where $\bar{\boldsymbol{x}}$ and $\boldsymbol{w}$ are mean and random component of prior information. Here, it is postulated that the observation $\boldsymbol{z}$ is expressed as a linear function of $\boldsymbol{x}$ and is contaminated with a Gaussian noise $\boldsymbol{v}$ as follows.

$$\boldsymbol{z} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{v} \tag{2}$$

$\boldsymbol{v}$ and $\boldsymbol{w}$ are Gaussian random variable vector with zero mean and their covariance matrices $\boldsymbol{R}$, $\boldsymbol{M}$. Best posterior estimate (MAP) and its covariance matrix are,

$$\hat{\boldsymbol{x}} = \bar{\boldsymbol{x}} + \boldsymbol{P}\boldsymbol{H}^T\boldsymbol{R}^{-1}(\boldsymbol{z} - \boldsymbol{H}\bar{\boldsymbol{x}}) \tag{3}$$

$$\boldsymbol{P} = (\boldsymbol{H}^T\boldsymbol{R}^{-1}\boldsymbol{H} + \boldsymbol{M}^{-1})^{-1} \tag{4}$$

Kriging is a probabilistic interpolation method in a Gaussian random field (e.g. Christakos 1992; Cressie 1991), and is derived as a special case of above mentioned linear inverse problem (Hoshiya and Yoshida, 1996). Assume that the observation vector $\boldsymbol{z}$ and parameters $\boldsymbol{x}$ are the same physical parameters at discrete spatial points in the Gaussian random field.

$$\boldsymbol{x}^T = \left\{\boldsymbol{x}_1^T, \boldsymbol{x}_2^T\right\} \tag{5}$$

where $\boldsymbol{x}_1$ denotes variables at observation site; $\boldsymbol{x}_2$ denotes parameters at the region to be estimated. The observation equation Eq.(2) becomes,

$$\boldsymbol{z} = \boldsymbol{H}\boldsymbol{x} = \begin{bmatrix}\boldsymbol{I} & \boldsymbol{0}\end{bmatrix}\begin{Bmatrix}\boldsymbol{x}_1\\ \boldsymbol{x}_2\end{Bmatrix} + \boldsymbol{v} \tag{6}$$

where $\boldsymbol{I}$ denotes unit matrix; $\boldsymbol{0}$ denotes zero matrix.

By substituting Eq.(5) , (6) into Eq.(3), (4), we have,

$$\begin{Bmatrix}\boldsymbol{x}_1\\ \boldsymbol{x}_2\end{Bmatrix} = \begin{Bmatrix}\bar{\boldsymbol{x}}_1\\ \bar{\boldsymbol{x}}_2\end{Bmatrix} + \begin{bmatrix}\boldsymbol{M}_{11}\\ \boldsymbol{M}_{12}^T\end{bmatrix}\left[\boldsymbol{M}_{11} + \boldsymbol{R}\right]^{-1}\left\{\boldsymbol{z} - \bar{\boldsymbol{x}}_1\right\} \tag{7}$$

$$\boldsymbol{P} = \begin{bmatrix}\boldsymbol{P}_{11} & \boldsymbol{P}_{12}\\ \boldsymbol{P}_{12}^T & \boldsymbol{P}_{22}\end{bmatrix} \tag{8}$$

$$\boldsymbol{P}_{ij} = E[\boldsymbol{M}_{ij} - \boldsymbol{M}_{i1}(\boldsymbol{M}_{11} + \boldsymbol{R})^{-1}\boldsymbol{M}_{1j}] \tag{9}$$

It is noted that Prior covariance matrix $\boldsymbol{M}$ is separated into $\boldsymbol{M}_{11}$, $\boldsymbol{M}_{12}$, $\boldsymbol{M}_{21}$, $\boldsymbol{M}_{22}$ corresponding to $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$,

Prior covariance matrix $\boldsymbol{M}$ is formulated often based on auto-correlation function. Several types of auto-correlation function are proposed. In this paper, the following equation is used.

$$R(d_1, d_2, d_3) = \sigma^2 \exp\left[-\left\{\left(\frac{d_1}{a_1}\right)^2 + \left(\frac{d_2}{a_2}\right)^2 + \left(\frac{d_3}{a_3}\right)^2\right\}\right] \tag{10}$$

where, d_1, d_2, d_3 stand for a distance, a_1, a_2, a_3 stand for an auto-correlation distance in each direction in three dimensional space; σ^2 is variance of the field.

2.2. Quantification of VoI

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.

i) Decision error type 1
Judge $x< x_0$ when true $x> x_0$ (e.g., to judge that soil is not contaminated when it is contaminated actually)

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

The probabilities of decision error type 1, 2 are denoted as P_1, P_2 (=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.

$$J = \sum_i L_i = \sum_i \min\left(C_1 P_{1,i},\quad C_2(1-P_{1,i})\right) \qquad (11)$$

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

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 which is model as Gaussian with mean=3, standard deviation=0.4. It is also assumed that penalties of the error type 1, 2 , C_1, C_2, are 10, 2 respectively. If the estimator is judged to be less than the threshold value, the probability of error is 0.5, and its risk is $10\times 0.5 = 5$. If the estimator is judged to be larger than the threshold value, the probability of error is also 0.5, and its risk is $2\times 0.5 = 1$. 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. Figure 1 shows the risk we should take when the mean is 0 to 4. and its standard deviation is 0.4.

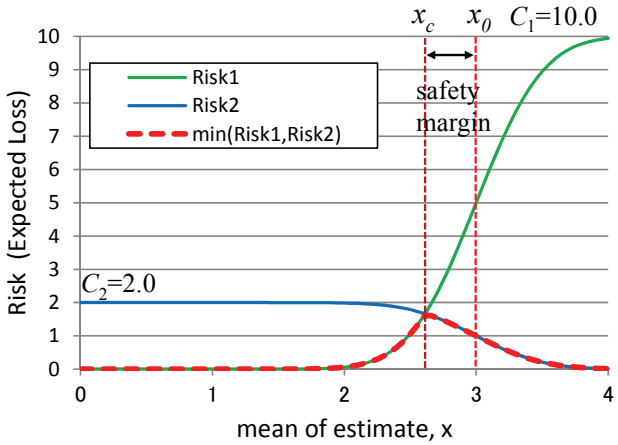


Figure 1: Risk of decision error, type 1 and 2, and mean of estimator, standard deviation of estimator=0.4, penalties C_1=10, C_2=2

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 take risk 1 when x is less than x_c, vice versa. 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.

In general, it is difficult to compute VoI so that MC approach is proposed (Pozzi and Kiureghian 2012; Straub 2013). VoI can be, however, computed easily in updating of Gaussian random field, i.e., Kriging, as described below. It is assumed that observation data at new locations are obtained at each observation step.

$$\boldsymbol{Z}^{k^T} = \left\{\boldsymbol{z}^{1^T}, \boldsymbol{z}^{2^T}, \cdots, \boldsymbol{z}^{k^T}\right\} \qquad (12)$$

where $\boldsymbol{z}^k$, $\boldsymbol{Z}^k$ represent observation data at step k and up to step k. Mean vectors at three types of places are obtained by referring Eq.(7),

$$\begin{Bmatrix} \bar{\boldsymbol{x}}_1^k \\ \bar{\boldsymbol{x}}_2^k \\ \bar{\boldsymbol{x}}_3^k \end{Bmatrix} = \begin{Bmatrix} \bar{\boldsymbol{x}}_1^0 \\ \bar{\boldsymbol{x}}_2^0 \\ \bar{\boldsymbol{x}}_3^0 \end{Bmatrix} + \begin{bmatrix} \boldsymbol{M}_{11}^0 \\ \boldsymbol{M}_{12}^{0\,T} \\ \boldsymbol{M}_{13}^{0\,T} \end{bmatrix} \left[\boldsymbol{M}_{11}^0 + \boldsymbol{R}_1^k\right]^{-1} \left\{\boldsymbol{Z}^k - \bar{\boldsymbol{x}}_1^0\right\} \tag{13}$$

where $\bar{\boldsymbol{x}}_1^k$ represents a mean vector at places where the observation $\boldsymbol{Z}^k$ is given; $\bar{\boldsymbol{x}}_2^k$ is a mean vector at places where new observation $\boldsymbol{z}^k$ will be given; $\bar{\boldsymbol{x}}_3^k$ presents a mean vector at area where decision error risk is evaluated. Their covariance matrices are given as:

$$\boldsymbol{M}_{ij}^k = \boldsymbol{M}_{ij}^0 - \boldsymbol{M}_{1i}^{0\,T} (\boldsymbol{M}_{11}^0 + \boldsymbol{R}_1^k)^{-1} \boldsymbol{M}_{1j}^0 \tag{14}$$

It is noted that locations of $\boldsymbol{x}_2$ are those of observation points at k+1 step, and the observation data is not obtained yet. As mentioned above, the penalty is imposed on false decision-making. The risk can be evaluated from the product of probability of false decision making and the penalty.

$$L(\bar{x}_{3,i}^k, \sigma_{3,i}^k) = \min\left(C_1 P_{1,i}, \quad C_2(1 - P_{1,i})\right) \tag{15}$$

where, $P_{1,i} = \Phi(\beta_i), \qquad \beta_i = \dfrac{\bar{x}_{3,i}^k - x_o}{\sigma_{3,i}^k}$

Φ is the standard Normal (Gaussian) cumulative distribution function; $\sigma_{3,i}^k$ is standard deviation of $x_{3,i}^k$ which can be obtained from diagonal component of covariance matrix $\boldsymbol{M}_{33}{}^k$ shown in Eq.(14). The total risk at the decision making area is given by:

$$J^k = \sum_i L(\bar{x}_{3,i}^k, \sigma_{3,i}^k) \tag{16}$$

The decision error risk is reduced by the new information $\boldsymbol{z}^{k+1}$. After we obtained observation vector $\boldsymbol{z}^{k+1}$, the mean and covariance matrix of $\boldsymbol{x}_3$ is updated as:

$$\bar{\boldsymbol{x}}_3^{k+1} = \bar{\boldsymbol{x}}_3^k + \boldsymbol{M}_{23}^{k\,T} \left[\boldsymbol{M}_{22}^k + \boldsymbol{R}_2^{k+1}\right]^{-1} \left\{\boldsymbol{z}^{k+1} - \bar{\boldsymbol{x}}_2^k\right\} \tag{17}$$

$$\boldsymbol{M}_{33}^{k+1} = \boldsymbol{M}_{33}^k - \boldsymbol{M}_{23}^{k\,T} (\boldsymbol{M}_{22}^k + \boldsymbol{R}_2^{k+1})^{-1} \boldsymbol{M}_{23}^k \tag{18}$$

Naturally value (number) of the new information $\boldsymbol{z}^{k+1}$ is not given yet. Therefore $\boldsymbol{x}_2{}^k$ instead of $\boldsymbol{z}^{k+1}$ is used in Eq.(17).

The expectancy of risk reduction is defined as VoI.

$$\text{VoI} = E[J^{k+1} - J^k] = E[J^{k+1}] - J^k \tag{19}$$

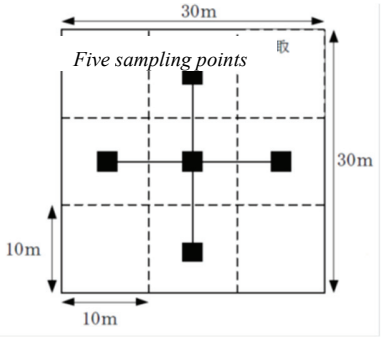
The expectancy of risk considering observation data in next step $\boldsymbol{z}^{k+1}$ is

$$E[J^{k+1}] = \sum_i \int L(\bar{x}_{3,i}^{k+1}, \sigma_{3,i}^{k+1}) p(\boldsymbol{x}_2^k) d\boldsymbol{x}_2^k \tag{20}$$

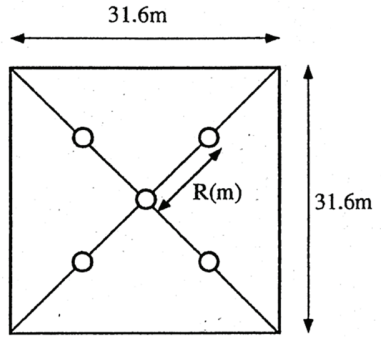
Integration with respect to $\boldsymbol{x}_2{}^k$ is required, but it cannot be performed analytically. When dimension of $\boldsymbol{x}_2{}^k$ is high, numerical integration is not easy to implement. Thanks to reproductive property of Gaussian, the numerical integration can be always reduced to one-dimensional numerical integration. Consequently VoI can be calculated easily even if the dimension of $\boldsymbol{z}^{k+1}$ (the number of additional observation points) is large, e.g., more than 10.

2.3. Optimization of VoI with respect to location of new observation

When the dimension of vector $\boldsymbol{z}^{k+1}$ is low, it is not difficult to optimize the location of new observation. The optimal locations can be determined by evaluating VoI at every possible combination of locations. It is, however, difficult to evaluate them due to "curse of dimensionality" when the dimension of the vector $\boldsymbol{z}^{k+1}$ 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 but efficient method for optimization with regard to real number variables.



(1) *Ministry of Environment, Japan*



(2) *Yoneda et al.(1999)*

Figure 2: Optimal placement of sampling

3. TWO DIMENSIONAL OPTIMAL PLACEMENT

3.1. Sampling planning for soil contamination

Many industrial sites can be potentially contaminated as a result of industrial development. Contaminated sites are found at many places such as motor workshops, petrol stations, fuel depots, railway yards, landfills and industrial sites. They are likely to pose an immediate or long term hazard to human health or the environment. Measures such as contamination remediation should be implemented at land which does not meet the designation standard based on the results of a soil contamination investigation.

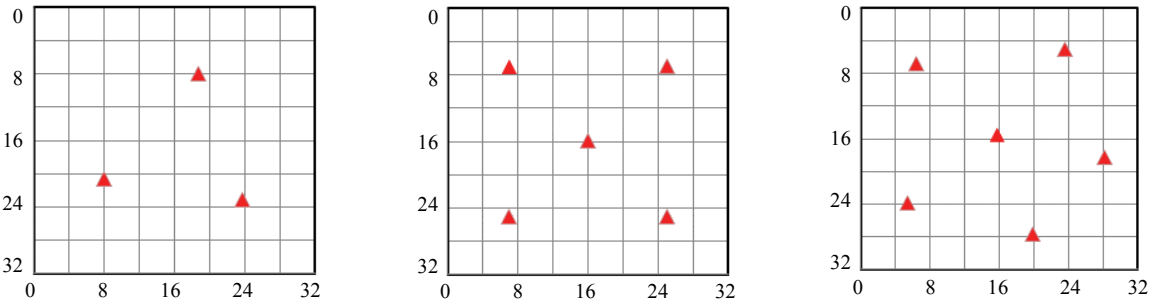
The Guidelines by Ministry of environment, Japan (2012) indicates detailed investigation scheme for the identification of contamination, in which a basic placement of sampling (sampling grid) is introduced for square area of about 1000 m^2 as shown in Figure 2(1). Yoneda (1999) studied the optimal location for soil contamination investigation in terms of uncertainty minimization. The obtained placement is shown in Figure 2(2). The placement is basically same as Figure 2(1), which is rotated by 45 degree at the center of the area.

3.2. Optimal placement without existing observation

Optimal locations are evaluated in terms of VoI under the assumption of Gaussian random field with respect to a contamination parameter. Its mean, standard deviation and autocorrelation distance are assumed to be 1.0, 0.5 and 15m respectively. The threshold value (critical value) for contamination x_0 is 2, which means that the area with larger than 2 is considered to be contaminated. The standard deviation of observation error is 0.1. The penalties C_1 and C_2 are 10 and 2.

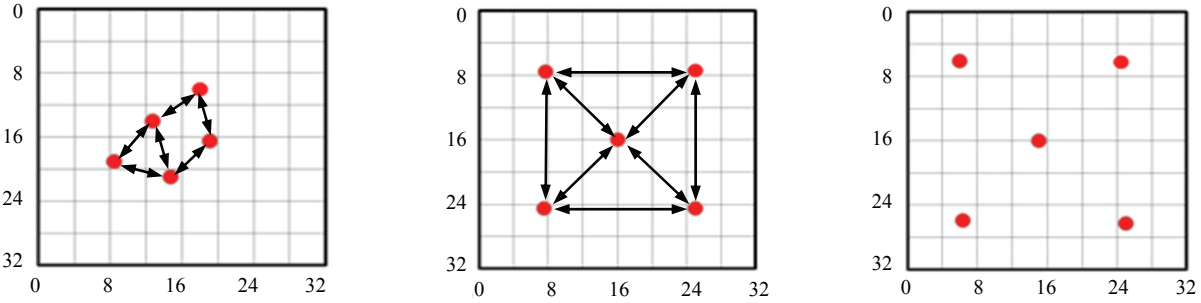
Figure 3 shows optimal placements without existing observation. They are obtained by minimization of VoI with respect to observation locations when the number of observation point is 3, 5 and 6. PSO is one of the heuristic methods with random numbers. The placements which are rotated from the placements shown in the figure are sometimes obtained depending on employed random numbers. The placement of 5 points is almost similar to Figure 2(2). It suggests that VoI-based method gives almost same optimal placement as uncertainty-based method in the case without existing observation.

Autocorrelation distance has a large influence on the optimal placement. Figure 4 shows the optimal placement when the autocorrelation distances are 3, 12 and 20 m. When the autocorrelation distance is large, the distance between the observation locations is also large. The shape of observation points is trapezoidal when the autocorrelation distance



(1) 3 *observation points* (2) 5 *observation points* (3) *6 observation points*

Figure 3 Optimal placement of observation points without existing observation



(1) d =3 m (2) d =12 m (3) d = 20m

Figure 4: Optimal placement and auto-correlation distance, five observation points (d=Auto-correlation distance)

d=3, while the shape is like five of dice when the distance d=12 or 20 m. The optimal placements are estimated with the autocorrelation distance from 3 to 20m. Figure 5 shows the relationship between the average length of arrows shown in the Figure 4 and the autocorrelation distance. They have clear proportional relations as shown in the figure. In the case of 5 points, the shape less than 10 m, which is trapezoidal, is different from one greater than 10m, which is like five of dice.

Total cost can be evaluated by sum of VoI and observation cost. The optimal placement and its VoI are calculated for the cases that the number of observation is one to seven. The distribution of total cost is shown in Figure 6 when the observation cost is assumed to be 8 or 5 per single observation point. The distributions of total cost show minimum points. The optimal number of observation is 3 when the observation

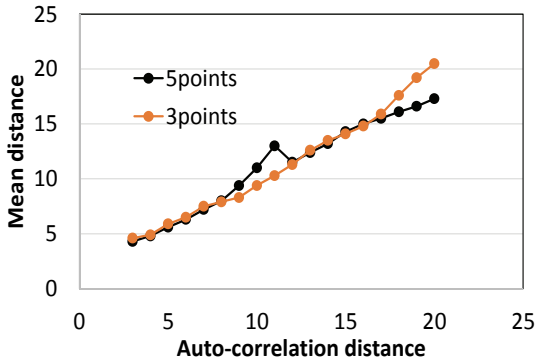
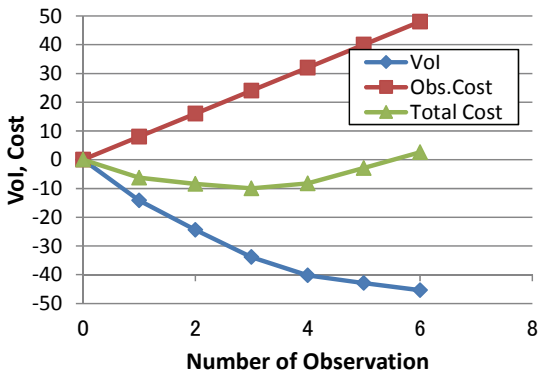


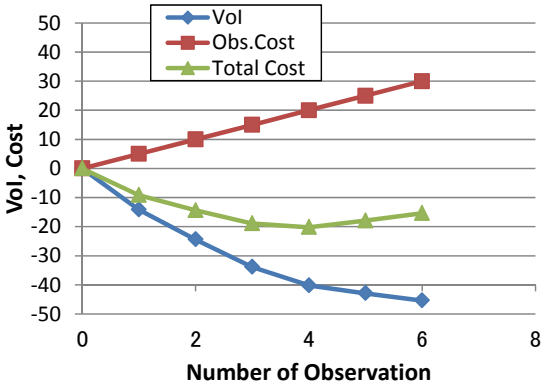
Figure:5 Mean distance among observation points and auto-correlation distance

cost is 8, while the optimal number is 4 when the cost is 5.

Basically the obtained placements agree with our intuition. It is not difficult to determine the placement intuitively when existing observation is not exist. More important problem



(1) *Observation* cost = 8



(2) *Observation cost* = 5

Figure 6: Optimal number of observation points

setting is optimal additional placement with existing observation data as discussed below.

3.3. Optimal placement with existing observation

It is assumed that there are three existing sampling, which are denoted by circles in Figure 7. The values obtained by the sampling are also shown in the figure. Three more additional sampling are performed for more reliable decision-making. We want to determine optimal locations for the additional sampling. Figure 7 shows obtained optimal locations for the additional sampling in terms of VoI depending on assumed autocorrelation distance, 7, 10 and 15m. The area near the sampling point of large value has higher possibility for remediation so that the additional sampling locations near the point are selected. It is noted that remediation measure is assumed to be taken when the value is larger than 2. When the autocorrelation distance is large, the distance between the additional observation points are large. This trend is similar to the cases without existing observation.

Figure 8 shows optimal placement depending on the values obtained at existing observation points. The number of additional observation is 2. Three cases are considered as to the values of existing observation. The optimal placements are determined depending on the observed values.

4. CONCLUSIONS

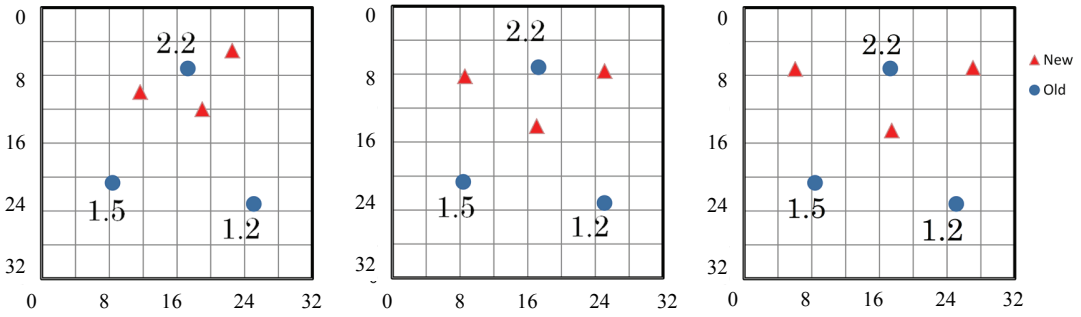
A Method for determining optimal observation placement in Gaussian random field is formulated based on Value of Information (VoI). Parameter study with respect to autocorrelation distance assumed in the Gaussian field indicates that when the autocorrelation distance is large, the distance between the observation points are large. This trend is similar to the both cases with or without existing observation. It is also shown that the shape of the optimal placement sometimes changes depending on the autocorrelation distance. Optimal number of observation is also evaluated based on total cost which is sum of observation cost and VoI. The proposed method can show that the optimal number is small when the observation cost is high. This is only to be expected result but the method can provide the optimal number quantitatively.

5. REFERENCES

Christakos, G. 1992. Random Field Models in Earth Sciences, Academic Press Inc.

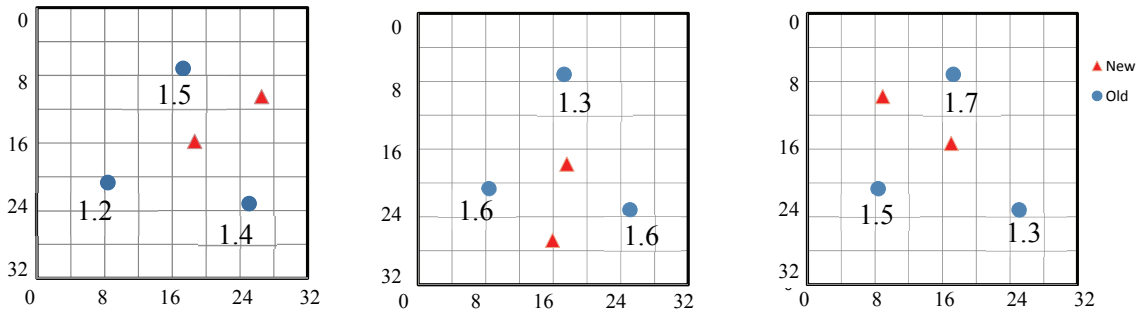
Cressie, N., 1991. Statistics for Spatial Data, John Wiley & Sons.

Honjo, Y. & Kudo, N. 1999. On inverse analysis and observa-tion scheme for ground deformation analysis based on in-formation entropy, Proc. 8th ICASP Conference, Applica-tion of Statistics and Probability, Balkema, 387-394.



(1) *d*=7m (2) *d*=10m (3) *d*=15m

Figure 7: Optimal additional sampling points. d=auto-correlation distance, abstract



(1) Case 1 (2) Case 2 (3) Case 3

Figure 8: Optimal placement of two additional observation points with variation of observed values

Hoshiya, M. & Yoshida, I. 1996. Identification of Conditional Stochastic Gaussian Field, Jour. of EM, ASCE, 122(2), 101-108.

Hoshiya, M. & Yoshida, I. 1998. Process Noise and Optimum Observation in Conditional Stochastic Fields, Jour. of EM, ASCE, 124(12), 1325-1330.

Kennedy, J. & Eberhart, R. 1995. Particle swarm optimization, Proc. of IEEE Int. Conf. on Neural Networks, Vol.4, 1942–1948.

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 Engineer-ing, 217-226.

Pozzi M. & Der Kiureghian A. 2012. Assessing the Value of Alternative Bridge Health Monitoring Systems, 6th Interna-tional Conference on Bridge Maintenance, Safety and Management, IABMAS: CRC Press.

Raiffa, H. & Schlaifer, R. 1961. Applied statistical decision theory. Boston: Clinton Press, Inc.

Straub, D. & Faber, M.H. 2005. Risk based inspection planning for structural systems, Structural Safety, 27, 335-355.

Straub, D. 2013. Value of Information Analysis with Structural Reliability Methods, Structural Safety, special issue in the honor of Prof. Wilson Tang

Sun, N-Z. 1994. Inverse problems in groundwater modelling, Kluwer Academic Publishers

Wu, S., Beck, J.L. & Heaton, T.H. 2013. ePAD: Earthquake Probability-Based Automated Decision-Making Frame-work for Earthquake Early Warning, Computer-Aided Civil and Infrastructure Engineering, 28(10), 737–752.

Yoneda, M., Morisawa, S. and Nishimura, R. 1999. Optimal Allocation of Sampling Points in the 5-Point Mixture Method for a Survey of Soil Contamination, JSCE, Journal of Environmental Systems and Engineering, 51-58. (In Japanese)