Mathematical Geology, Vol. 37, No. 1, January 2005 (© 2005)

An Interactive Sampling Strategy Based on Information Analysis and Ordinary Kriging for Locating Hot Spot Regions[1]

Shyan-Shu Shieh,[2] Ji-Zheng Chu,[3] and Shi-Shang Jang[4]

This study proposes an interactive sampling strategy for locating the hot spot or maximum regions of a concerned attribute in a given area of survey. In the proposed strategy, information analysis is performed based on the ordinary kriging from the existing sample data to suggest a new batch of samples under the criterion of the highest information free energy. The information free energy (F) is a function of information energy (U) and information entropy (S) through $F = U - TS$, where T is information temperature and is used to coordinate the contribution of U and S to F. Information energy is the value of the concerned attribute, and information entropy is the transformed error variance of kriging and therefore measures the evenness and density of coverage of samples over the area under survey. At early sampling batches, information temperature is high and information entropy dominates the information free energy, and samples are suggested to give an even and dense enough coverage of the whole area under investigation. As samples accumulate, information temperature decreases to enlarge the contribution of information energy, and future samples are taken toward the locations with high attribute values. Two examples demonstrate the efficiency and effectiveness of the proposed sampling strategy in locating the hot spot regions of various fields: (1) a heavy metal contaminated site reproduced by modeling on 55 real field data; (2) a simulated two-dimensional field by the random phase volume (RPV) model. The results show that the proposed strategy, a robust interactive sampling procedure, is able to locate hot spot regions without compromising with the overall profile of an under-survey area.

KEY WORDS: information theory, intermediate modeling, multi-stage sampling, random phase volume.

INTRODUCTION

There are a large number of wastes dumping sites in the world, which were formed illegally and in great hurry or legally but without any record of their constitutions.

[1]Received 19 August 2003; accepted 27 July 2004.
[2]Department of Occupational Safety and Hygiene, Chang Jung University, Tainan.
[3]Department of Automation, Beijing University of Chemical Technology, Beijing.
[4]Chemical Engineering Department, National Tsing-Hua University, Hsin-Chu 30043; e-mail: ssjang@che.nthu.edu.tw

0882-8121/05/0100-0029/1 © 2005 International Association for Mathematical Geology

P1: KJT
JID: matg File: matg-sjny005-498746 AID: sjny005 MID: 498746 PIPS: 498746 November 10, 2004 23:44

Such large-scale dumpling sites of waste constitute a serious environmental problem and their evaluation through sampling is of great importance.

In surveying the distribution of a particular attribute in an area, one of the main concerns is where "hot spots", or maxima, or extremes are and what the attribute values are at hot spots. For instance, we hope to know not only the spreading extent of contamination, but also the highly polluted sub-regions in a suspected area. Both are equally important in determining the financial needs in cleaning up, in deciding the feasible treatment technology, and in assessing the risk of environmental exposure. The concerns are also very true in mining. Rich areas of a mine often decide the value of the whole field. In fact, the long and labor-consuming process of a survey needs the encouragement from finding hot spot regions. The distributing extent and the hot spot regions are two interrelating aspects of a survey. Only when all the main hot spot regions of the interested attributes are identified, an area can be regarded as thoroughly understood. On the other hand, without a sufficient coverage of samples over the investigated area, it is not possible to find all the hot spot regions.

Geostatistical methods map the spatial variability of attributes by interpolating. To a large extent, the number and distribution of samples determine the quality of mapping and the cost. Many researchers have since attempted to design efficient sampling schemes. In a discussion paper, Brus and de Gruijter (1977) showed that both the design-based sampling strategy stemmed from classical sampling theory, and the model-based sampling strategy developed in geostatistics, are valid for spatial sampling and estimation. They also noted that many factors determine the effectiveness and efficiency of a statistical approach for spatial sampling and estimation. Therefore, they developed a decision tree for selecting between the model-based and the design-based sampling strategies. Aiming at the full use of the prior information on soil variability and statistical knowledge on spatial sampling, Domburg, de Gruijter, and van Beek (1997) proposed a knowledge-based system using dynamic programming for designing efficient soil survey schemes. van Groenigen, Stein, and Zuurbier (1997) proposed an interactive sampling procedure to optimize environmental risk assessment. In their method, probability maps are made with indicator kriging from the existing spatial interpolation results, and such probability maps are then used to direct subsequent sampling.

Sampling design is a complex and constrained optimization problem. {Domburg,} de Gruijter, and van Beek (1997) pointed out that sampling schemes should be designed so that either the costs are minimized under certain mapping quality requirements, or the quality is maximized within a given budget. When the problem is limited to the optimal distribution of sample points in a given area, two kinds of partially conflicting sampling strategies can be classified aiming at optimal estimation of variogram parameters and at optimal spatial interpolation respectively (van Groenigen, Pieters, and Stein, 2000). As for sampling strategies aiming at optimal spatial interpolation, van Groenigen, Siderius, and Stein (1999)

introduced the extended spatial simulated annealing (SSA) method to optimize spatial sampling schemes through minimizing the kriging variance, which allows considering the effect of previous observations and boundaries. Their method was further extended to multivariate problems through an optimization criterion called weighted means of shortest distance (van Groenigen, Pieters, and Stein, 2000). In the methods of the above two papers, the criterion of optimization is basically a measure of distance or statistical distance between the sample points and the raster points representing the whole area of survey, although in multivariate problems, weights that are a measure of the attribute values have an effect on the final sampling scheme.

As we have noticed at the beginning that hot spot regions have special meanings not only for themselves but also for the estimation of the whole profile, missing a main hot spot region often means gross error in the estimated profile. In order to make a more efficient sampling, Watson and Barnes (1995) established three problem-dependent meanings for engineering extremes and translated them into formal geostatistical/model-based criteria for designing infill sample networks. Sasena, Papalambros, and Goovaerts (2000) made a systematic comparison between the three criteria of Watson and Barnes and the so-called efficient global optimization (EGO) algorithm (Jones, Schonlau, and Welch, 1998) which uses a generalized expected improvement function to decide new sample points, and concluded that none of the criteria is superior in all respects.

Generally speaking, any optimization procedure is a trial and error process of making decisions on where a trial should be taken and performing the trial at the selected conditions or locations. In the case of sampling design where prior knowledge about the system is unavailable or scarce, expensive and time-consuming sampling and sample analysis activities are necessary. Efforts on designing efficient sampling schemes can be justified by the saving on laboring and laboratory expense. In our previous study (Chen and others, 1998), an artificial neural network (ANN) was employed as a universal modeling tool (a meta-model) for correlating a performance index with operation variables from existing experimental data, and information analysis was then performed on the established model to suggest conditions at which future experiments are needed. The information analysis maximizes the information free energy (F), which is a function of information energy (U) and information entropy (S). The information energy, or the performance index, is usually taken as the value of the interested attribute. On the other hand, the information entropy is a measure of the uniformity extent either in the distribution of the experimental conditions in the space of operation variables for experimental design or in the location of sample points in the spatial coordinates for sampling design. Another quantity called information temperature (T) is employed to coordinate the contribution of information energy and information entropy to the information free energy through $F = U - TS$, which is an analogue to the concept of free energy in thermodynamics.

At early sampling stages, information temperature is high and information entropy dominates the information free energy, which suggests an even coverage of samples on an investigated area. As a result, a rough, but somehow accurate enough profile is being evolved to reveal the main features of an investigated site. The information temperature decreases with the increase of samples and information energy becomes more and more important to the information free energy. The locations of sampling are therefore toward hot spot regions.

The criterion in the above information analysis procedure (Chen and others, 1998) is similar to the criterion for locating the regional extreme of Watson and Barnes (1995) and to the generalized expected improvement function (with $g = 1$, see Sasena, Papalambros, and Goovaerts, 2000). However, these criteria have different roots in their derivation. While the criteria of Barnes and Watson and the generalized expected improvement function are derived from statistics, the criterion of the proposed information analysis comes from an analogue to thermodynamic free energy. Being inspired by the successful application of information analysis to various optimization processes (Chen and others, 1998, 1999; Lin and Jang, 1998; Chu and others, 2003; Yen, Wong, and Jang, 2003), we propose an interactive sampling (IS) strategy based on the above procedure.

In the following context of this paper, system definition and a brief introduction to the OK-estimator are firstly presented in section on sample system definition and ordinary kriging estimates. Section on interactive sampling strategy based on information analysis is dedicated to the development of the proposed interactive sampling strategy. Section on two demonstration examples is used to show the results of applying the proposed sampling scheme to the survey of two simulated fields. The last section is for conclusion.

SAMPLE SYSTEM DEFINITION AND ORDINARY KRIGING ESTIMATES

For a space Ω containing

$$V = \{(y_i, x_i) \mid \quad i = 1, 2, \ldots, n\} \tag{1}$$

the ordinary kriging estimates the function value at a point x_0 in Ω by a weighted linear combination as

$$\tilde{y}_0 = \sum_{i=1}^{n} w_i y_i \tag{2}$$

where x is the position coordinate of a point, y and $\tilde{y}$ are the measured and estimated function values, and $w_1, w_2, \ldots, w_n$ are the kriging weight determined

by the following formula

$$W = C^{-1}D \tag{3}$$

with

$$W = [w_1, w_2, \ldots, w_n, \mu]^T \tag{4}$$

$$C = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1n} & 1 \\ c_{21} & c_{22} & \cdots & c_{2n} & 1 \\ \vdots & \vdots & \ldots & \vdots & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nn} & 1 \\ 1 & 1 & \cdots & 1 & 0 \end{bmatrix} \tag{5}$$

$$D = [c_{10}, c_{20}, \ldots, c_{n0}, 1]^{\mathrm{T}} \tag{6}$$

where μ is a Lagrange multiplier derived from the kriging algorithm, and c_{ij} is a statistical distance between points i and j. The relation between c_{ij} to its corresponding geometric distance h_{ij} is called the covariance model. The most popularly used model of covariance is the exponential model and has the following form

$$c = \begin{cases} c_0 + c_1, & \text{if } |h| = 0 \\ c_1 \exp\left(\frac{-3|h|}{a}\right), & \text{if } |h| > 0 \end{cases} \tag{7}$$

where c_0, c_1, and a are parameters to describe the continuity in the space to be investigated, and should be chosen carefully according to sample data as well as experience. c_0 is called the nugget effect, $c_0 + c_1$ the sill, and a the range.

Isaaks and Srivastava (1989) discussed the effect of these parameters on the estimates in detail. In the case of anisotropy, special technique is required to get a reduced covariance model by combining all covariance models recognized for different directions. van Groenigen (2000) explained the influence of variogram parameters on optimal sampling schemes for mapping by kriging. However, our experience showed that the sensitivity of the sampling scheme resulted from the procedure proposed in this study to these parameters is not high. Therefore, we treat the ordinary kriging estimator in this study simply as a substitute to the artificial neural network in the paper of Chen and others (1998), which serves as a meta-model tool, and the parameters of the variograms are valued quite arbitrarily to be $c_0 = 0$, $c_1 = 10$, and $a = 10$ for both two case studies in section on two demonstration examples. Isotropy is also assumed in the case studies. It should be noted that the kriging estimator using a better-suited variogram model is meaningful in practical applications. The above variogram model may be very rough for the sampling spaces of the case

studies in this paper, and it just serves to construct a basis for demonstration and comparison.

INTERACTIVE SAMPLING STRATEGY BASED ON INFORMATION ANALYSIS

The objective of the proposed work is to find hot spot regions in an unknown field with the minimum number of sampling. To achieve the goal, we propose an iterative process of sampling instead of one-time sampling. The idea is to utilize partial (or prior) knowledge obtained from the existing sampling data to reduce the number of sampling data on uninterested sub-areas. We use the ordinary kriging estimator on the existing samples to obtain an up-to-date model as the partial knowledge which can provide some clues toward the next sampling locations to find hot spot regions. An up-to-date model, serving as partial knowledge, is also called a meta-model. We consider regions with smooth landscape or with low attribute values as uninterested area contrary to hot spot regions.

Besides reducing the number of sampling data via the clues from the existing sampling data, remotely untapped area is, on the contrary, worth our attention to take samples. To resolve the issues, we propose information analysis, which introduces information energy and information entropy. In the following theoretical derivation, we will show how to balance the choice of sampling locations on these two factors.

According to Shannon's definition (Shannon, 1948; Shannon and Weave, 1949) of information entropy for a variable X, which can randomly take values x from a set X, the information entropy of the set X is

$$S(x) = \sum_{x} p(x) \ln[p(x)] \tag{8}$$

where $p(x)$ is the probability of the event x occurring. If the variable X can only take a narrow range of values, $p(x)$ for these values is close to 1. For other values in X, $p(x)$ is close to 0. Therefore $S(x)$ is close to zero. If the variable X can take a lot of different values in X each time with a small $p(x)$, $S(x)$ will be a large negative number. Thus, information entropy is a measure of how random a variable is distributed. It decreases when the variable is more randomly distributed.

Let us apply the information entropy in this study to measure the evenness of sample locations among the whole surveyed area. Information entropy can be calculated from its definition by assuming some probability distribution of the concerned attribute (Haykin, 1999). According to the maximum entropy principle, the transferred error variance of kriging can be taken as information entropy when the ordinary kriging estimator is used as a meta-model.

Suppose that there already exist samples $V' \in \Omega$ and

$$V' = \{(y_i, x_i) | \quad i = 1, 2, \ldots, n - m\} \tag{9}$$

One of our problem is to find the optimum positions x_k ($k = n - m + 1, \ldots,$ n) at which m new samples will be taken. As a result, the overall estimates about the distribution in Ω will improve to the largest degree of possibility by the addition of these new samples $(x_{n-m+1}, y_{n-m+1}), \ldots, (x_n, y_n)$.

In solving the above problem, let us firstly review the derivation of the ordinary kriging (Isaaks and Srivastava, 1989). The ordinary kriging is called the best linear unbiased estimator (BLUE) because the linear weighted combination is used, the expected mean error is zero, and the variance of error is minimized. Given a set of samples, $V = V' \cup (y_{n-m+1}, x_{n-m+1}) \cdots \cup (y_n, x_n)$, the minimum error variance can be evaluated by

$$\sigma_R^2 = \sigma^2 - W^{\mathrm{T}} D \tag{10}$$

or equivalently

$$\sigma_R^2 = \sigma^2 - D^{\mathrm{T}} C^{-1} D \tag{11}$$

where σ^2 is variance of random variable $\tilde{y}_0$.

Apparently, σ_R^2 is a function of both the sample set (number and distribution of the sample points) and the position x_0 at which $\tilde{y}_0$ is evaluated. In the framework of kriging estimator, the unbiasedness is guaranteed by a constraint in the derivation, and the optimum position of the next sample points can be determined by the following minimization as done by van Groenigen, Siderius, and Stein (1999)

$$\min_{x_{n-m+1}, \ldots, x_n \in \Omega} \int_{\Omega} \sigma_R^2(x_0, x_{n-m+1}, \ldots, x_n)\, dx_0 \tag{12}$$

In the ordinary kriging algorithm, the predicted error at a point x_0 is a random variable $R = r \in (-\infty, \infty)$ with zero mean and a variance of σ_R^2. For such a random quantity, the maximum entropy principle (Haykin, 1999) says that it should follow a Gaussian distribution

$$f_R(r) = \frac{1}{\sqrt{2\pi}\sigma_R} \exp\left(-\frac{r^2}{2\sigma_R^2}\right) \tag{13}$$

and the corresponding information entropy is

$$S_R = \frac{1}{2}\left[1 + \log(2\pi\sigma_R^2)\right] \tag{14}$$

Equation (14) is the entropy of the prediction error for a particular point in space Ω, and the mean entropy of the whole space can be calculated as follows

$$S = \frac{1}{N_e} \sum_{i=1}^{N_e} S_{R,i} \tag{15}$$

where N_e is the number of points used to represent the whole space (also of the nodes of fine raster grid in the paper of van Groenigen, Siderius, and Stein, 1999).

The above criterion only considers the influence of statistical distance. If isotropic covariance models are used, the above criterion produces samples located uniformly in the space Ω. In accordance with our purpose to locate hot spot regions, it is wiser to populate more points at the maxima of the landscape, if there is some implication from the existing samples to show the landscape of merit. Namely, the ideal positions of the new samples should satisfy both the criterion of (12) and the following one:

$$\max_{x_{n-m+1},\ldots,x_n \in \Omega} \tilde{y}(x_{n-m+1}) + \cdots + \tilde{y}(x_n) \tag{16}$$

The criterion of (16) says that the new sample points should locate at the maximum of the landscape of merit. Since the two criteria of (12) and (16) generally conflict to each other, a coordination mechanism is necessary and can be built by defining a new quantity (Chen and others, 1998),

$$F = U - TS \tag{17}$$

where F is information free energy, $U = \tilde{y}(x_{n-m+1}) + \cdots + \tilde{y}(x_n)$is information energy, S is information entropy, and T is information temperature or temperature in short. Our problem now becomes $\max_{x_{n-m+1},\ldots,x_n \in \Omega} F$ for locating points at which F is maximized.

The temperature T in Equation (17) can be regarded as a compromising factor. At early sampling stages, T should be high, and the information entropy dominates to ensure a sufficiently uniform coverage of the whole space. With the increase of samples, the landscape of merit becomes clearer, and T is lowered to increase the effect of information energy, which will enable new samples to move toward the maximum of the landscape.

Some patterns exist for the decay of temperature in literature. The information temperature is a coordinating factor in nature and can be selected empirically. The magnitudes of U and S vary in different cases that have different units and scales. The choice of T then becomes difficult and depends on the value of U and S in each case. Therefore, we normalize U and S. A simple step function is used in this study for the information temperature $T = 10, 6, 4, 2, 0$ for batches 2, 3, 4, 5 and 6. After normalization, we rewrite Equation (17) as follows.

$$\bar{F} = \bar{U} - T\bar{S} \tag{18}$$

where $\bar{F}$, $\bar{U}$ and $\bar{S}$ are normalized infofrmation free energy, energy and entropy. The normalization can be done with the following equations

$$\bar{U} = \frac{U - U_{\min}}{U_{\max} - U_{\min}} \tag{19}$$

and

$$\bar{S} = -\frac{S - S_{\min}}{S_{\max} - S_{\min}} \tag{20}$$

together with

$$U_{\min} = \min(\tilde{y}_i \,|i = 1, 2, \ldots, N_e) \tag{21}$$

$$U_{\max} = \max(\tilde{y}_i \,|i = 1, 2, \ldots, N_e) \tag{22}$$

$$S_{\min} = \min(S_{R,i} \,|i = 1, 2, \ldots, N_e) \tag{23}$$

$$S_{\max} = \max(S_{R,i} \,|i = 1, 2, \ldots, N_e) \tag{24}$$

From the above analysis, the proposed interactive sampling strategy can be summarized in the following iterative sequence:

a. If no prior observations exist, samples are firstly taken in the area. In all the case studies of this paper, regular square grid scheme is employed to take the first batch of samples, which is necessary to start the proposed interactive sampling scheme.
b. The area to be investigated is specified together with the variogram model to be used, and nodes of fine raster grid for representation of the whole investigated area are established.
c. Give the number of samples (m) to be taken in the following batch and search for the locations of all the m samples through $\max_{x_{n-m+1},\ldots,x_n \in \Omega} \bar{F}$. Simulated annealing (Kirkpatrick, Gelatt, and Vecchi, 1983; van Groenigen, Siderius, and Stein, 1999) is used for the maximization.
d. Check the stopping criterion, which may be that the number of samples reaches a prescribed limit or that the suggested sample locations cluster closely.
e. If the stopping criterion is fulfilled, the whole procedure is terminated; otherwise, a batch of samples is taken at the locations determined in (iii), and go back to (ii).

At this point, it should be emphasized that the number of batches and the number of sample points in each batch should be scheduled depending on the available financial resources available, the time limitation, and other restrictions. After the numbers are arranged, the pattern of temperature decay is determined.

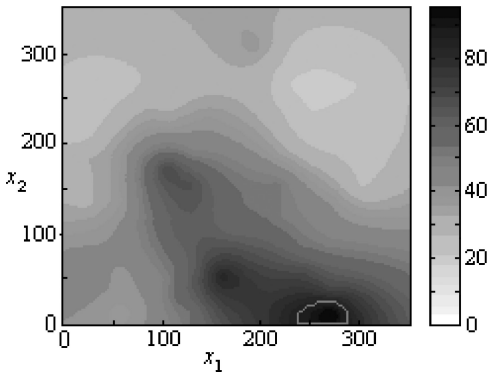

Figure 1. Attribute map of the HMC site and the hot spot region (attribute >86).

Both the numbers and the temperature pattern are crucial to the final results, and can be readjusted between batches or in a batch according to the judgment of an experienced operator. From this perspective, the algorithm above is a sampling philosophy to a large degree and also depends strongly on the experience of its users.

TWO DEMONSTRATION EXAMPLES

Case 1: A Heavy Metal Contaminated Site

Juang, Lee, and Chen (1996) surveyed a heavy metal contaminated site, approximately a 360 m by 360 m, in Taiwan by taking 55 field samples randomly. Although we never know the real contamination spatial profiles, we reproduce the site with the OK-estimator based on the 55 field data. Assuming the reproduction site is real, 100 by 100 grid points, totally 10,000, are estimated from the OK-estimator model as shown in Figure 1. The heavy metal contamination concentration ranges from 9.5 to 126.7. The contamination areas whose concentration is above 86 are considered as hot spot regions in this case. As a result, 79 out 10,000 grid points are deemed as in the hot spot region circled by the white-lined curved in Figure 1. From the range description part in Table 1 and the histogram in Figure 2, we can see more than 80% of points are in the uninterested area. Since we are concerned with the high attribute value region and the hot spot region, we divide the whole range into four sub-ranges according to the attribute value, not quartile of the total number. The number and mean of the sub-ranges, the total population and the hot spot region are listed in Table 1.

In this case, besides applying the proposed interactive sampling strategy to take samples from the above reproduced site, we also use the typical sampling

Table 1. Result Summary for the HMC Site

	Attribute range					Hot spot region
	0–25%	25–50%	50–75%	75–100%	0–100%	
Range description						
No. of points	5286	2892	1551	271	10000	79
Mean	29.09	47.21	65.91	83.53	41.51	90.27
Mean absolute errors of RG						
16 samples	2.32	3.32	5.03	18.05	3.45	25.7
25 samples	1.27	2.20	2.36	12.00	2.00	19.11
36 samples	0.97	1.62	2.76	9.29	1.63	14.41
49 samples	0.81	1.29	1.98	6.40	1.28	10.93
64 samples	0.53	0.90	1.41	4.94	0.90	9.34
Mean absolute errors of IS						
17 samples (Batch 2)	2.19	3.04	4.28	10.09	2.97	18.03
25 samples (Batch 3)	1.48	3.04	3.55	9.08	2.46	16.61
33 samples (Batch 4)	1.36	1.98	1.89	9.28	1.84	16.80
41 samples (Batch 5)	0.96	1.55	1.64	4.30	1.33	8.53
42 samples (Batch 6)	0.96	1.54	1.76	3.16	1.31	5.04

method, i.e., regular square-grid sampling for comparison. In the following text, IS denotes for interactive sampling while RG for regular square-grid sampling. The ordinary kriging estimator is employed to make models for both sampling methods.

For RG method, we take five times of sampling, i.e., 16 ($=4^2$), 25, 36, 49, 64 data each. For the convenience of comparison, the OK estimating results of the five-time samplings are also incorporated in Table 1. The mean absolute errors (MAE in short) in the columns of the sub-ranges, the total population, and the hot

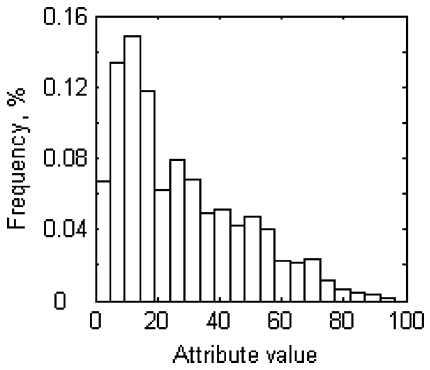

Figure 2. Histogram of 10,000 points in the HMC site.

spot region decrease as sampling size increases. It is natural that large sample size gives more accurate estimating for the smooth landscape.

For IS method, the first batch starts with nine square-grid samples. The OK estimator makes a model based on these nine samples. We, then, conduct information analysis that evaluates information energy via the nine-sample model and information entropy by Equation (14). We take eight samples for the next batch and each of the following batches. It is arguable of how many samples to take for each batch. However, the issue is beyond the scope of this study and we choose eight arbitrarily. The locations of the best eight samples giving the largest information free energy are subsequently determined.

With eight samples in Batch 2 and nine in Batch 1, the OK estimator is used to model on these 17 points. Repeating the interactive process, we take eight samples in Batch 3, 4 and 5. For Batch 6, all the best eight points are very close to each other. We decide to take just one of them and conclude the iteration process converging. In the consecutive sampling process of six batches, 42 samples are taken.

Scrutinizing on Table 1 that shows the results of comparing RG and IS methods, we can summarize the following points. The MAEs of the total population and the hot spot region decreases as the interactive sampling process proceeds. It is worth mentioning that at the beginning, through Batches 2–4, the lower attribute ranges (0–75%) improve dramatically while the 75–100% range, and the hot spot region are stagnant. The opposite phenomena, the 75–100% range and the hot region improving but the lower attribute ranges not, appear through Batches 4–6. It explains that at the beginning, the high temperature makes information entropy, S, more dominant and information energy, U, less in Equation (17). The controlling S in these stages favors even spreading of sampling points. At the final stages of sampling process, with decreasing temperature, U becomes dominating and favors sampling around the hot spot region. Even with the addition of just one sample from Batches 5 to 6, the MAE of the 75–100% range and the hot spot region decrease impressively because this sample hit the jackpot.

Figure 3 shows the detailed process of sampling-location determination. In the left-hand site, denoted as (a) in the figure, the samples of the current batch in the IS process are located with the background showing the contour map based on the 55 field data which is presumed the real site. In the right-hand site, denoted as (b) in the figure, the samples accumulated up to the present batch are located with the background showing the contour map of the up-to-present-batch model. In the first three batches, the locations of samples are symmetrical because of the domination of information entropy S. Comparing (a) and (b) in the first batch of nine samples, the figure shows that the model has formed a rough, but somehow accurate contour. Therefore, the chosen samples of Batch 2 are located symmetrically toward to the bottom of site. Starting from Batch 4, most of the chosen samples move toward the hot spot region with a few locating remotely.

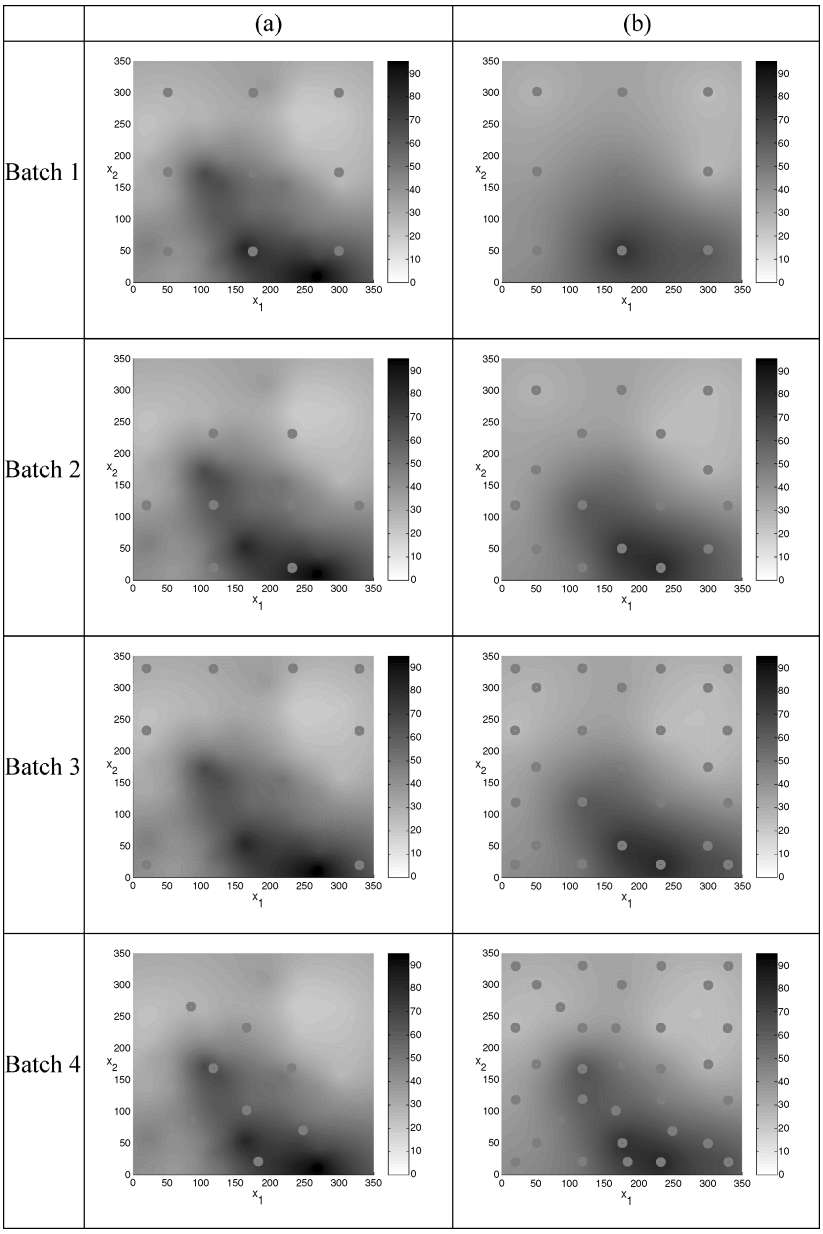

Figure 3. IS process for the HMC site: (a) contour map based on the 55 field data and samples taken at the current batch; (b) contour map of the current model based on the accumulation samples up to the current batch.

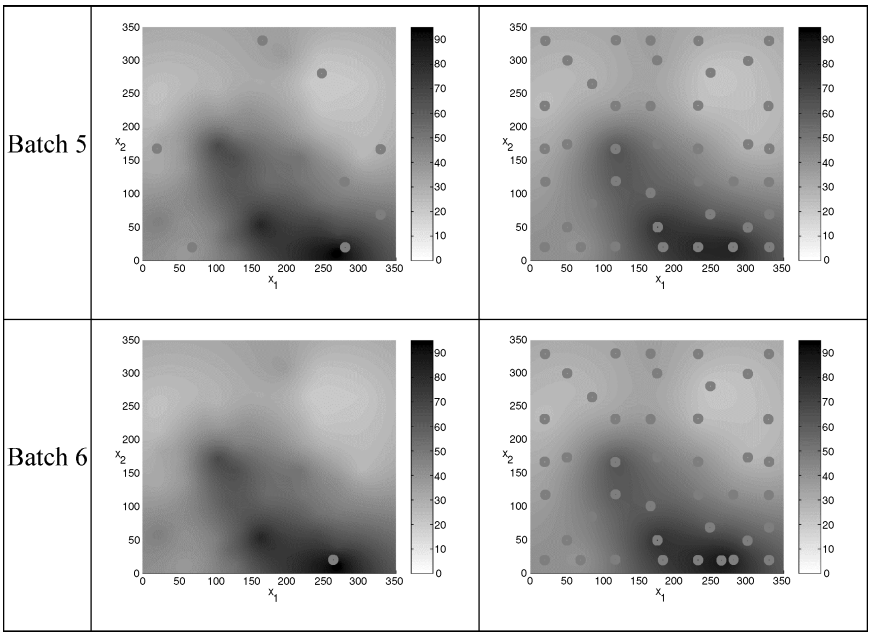

Figure 3. Continued.

The total number of samples taken in the IS method is 42 which is between 36 and 49 samples of the RG method. The MAEs of the lower attribute ranges (0–75%) are comparable between these two methods, but the MAEs of the 75–100% range and the hot spot region are much better in IS i.e., 3.16 and 5.04, than in RG i.e., 4.94 and 9.34 for 64-sample model. In this case, the promising results of predicting hot spot region in the IS method justify the effectiveness of information analysis on selecting sampling locations.

Case 2: A Two-Dimensional Field from the RPV Model

The last case reproduced from the OK simulation model based on the 55 field data. During simulating, we assume zero nugget effect, i.e., $c_0 = 0$ in Equation (7). Without any clue about the real contamination profile, we presume it is smooth. In a (especially illegal) waste-dumping site, wastes are dumped batch by batch from different sources. As a result, it is expected that a site contains various regions with different magnitudes of contamination and discontinuity exists along various regions. In this case, we generate a simulation case with very rugged landscape to see if the IS method can locate hot spot regions or not. However, during modeling

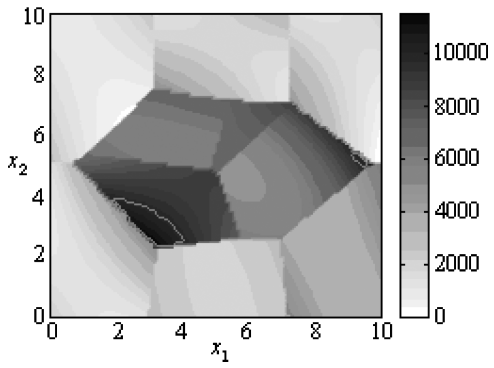

Figure 4. Attribute map of the 2D-RPV field and the two hot spot regions (attribute >10,000).

on sample data no matter of using the RG or IS method, we still assume to deal with smooth landscape pretending to know nothing ahead. In other words, we also assume zero nugget effect during modeling in this case.

We apply the random phase volume model (RPV in short) to generate a rugged landscape case. Falcioni and Deem (2000) have a detailed description of the RPV model in their article. The RPV model can generate a number of discontinuous regions (called phase in the original paper) in a multi-dimension space with various magnitudes and changing rates, e.g., sharp or slow rising and declining along landscape, among different regions.

Figure 4 is a two-dimensional field of ten phases produced from the RPV model (the functional form of the RPV model and the parameters used to generate Fig. 4 will be provided on request). For simplicity, we will name this simulation site as the 2D-RPV field. Figure 5 shows the histogram of 10,000 uniformly square-grid points. Since we are mainly concerned with the regions of high attribute values, we identify all the 150 points whose attribute values are over 10,000. Two hot spot regions, confining the 150 points, are enveloped as shown in Figure 4.

Figure 6 shows the detailed process of sampling location determination by the IS method. The process starts with nine samples in the first batch, and eight samples are taken for each of the following batches. Similar to Case 1, the locations of samples taken in Batch 2 and 3 are symmetrical due to the domination of information entropy factor. Samples suggested by IS in Batch 5 cluster around the neighborhood of the maximum point. Samples in the last batch converge to the maximum point predicted by the OK-estimator based on all the existing samples at the end of Batch 5. It is the result of information temperature's dropping. We take one sample in this batch and terminate the iterative sampling process.

Table 2. Result Summary for the 2D-RPV Field

	Attribute range					Hot spot region
	0–25%	25–50%	50–75%	75–100%	0–100%	
Range description						
No. of points	5150	2327	1801	722	10000	150
Mean	1692	4131	6887	9411	3753	10544
Mean absolute errors of RG						
16 samples	560	729	2010	3445	1069	5285
25 samples	1164	1112	623	1077	1048	1789
36 samples	485	655	1727	2091	864	3630
49 samples	505	847	1001	1690	760	3397
64 samples	396	565	861	1727	615	2897
Mean absolute errors of IS						
17 samples (Batch 2)	966	953	839	1795	1000	2765
25 samples (Batch 3)	785	913	872	1861	908	3003
33 samples (Batch 4)	849	958	781	1155	884	1553
41 samples (Batch 5)	860	997	776	976	886	1144
42 samples (Batch 6)	860	997	776	976	886	1144

As contrasted to the deterministic manner of IS in locating hot spot regions batch by batch, the RG scheme finds hot spot regions by chance when samples are not dense enough. This is clear by observing the MAEs of hot spot regions in Table 2. For the 2D-RPV field, RG of 25 points gives the lowest MAE for the two hot spot regions, whereas RG of 36, 49 and 64 points present much higher MAEs for the hot spot regions. The fact that one of 25 RG points hit the hot spot regions

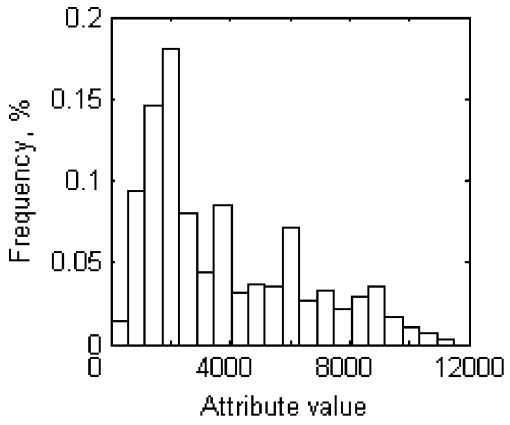

Figure 5. Histogram of 10,000 points in the 2D-RPV field.

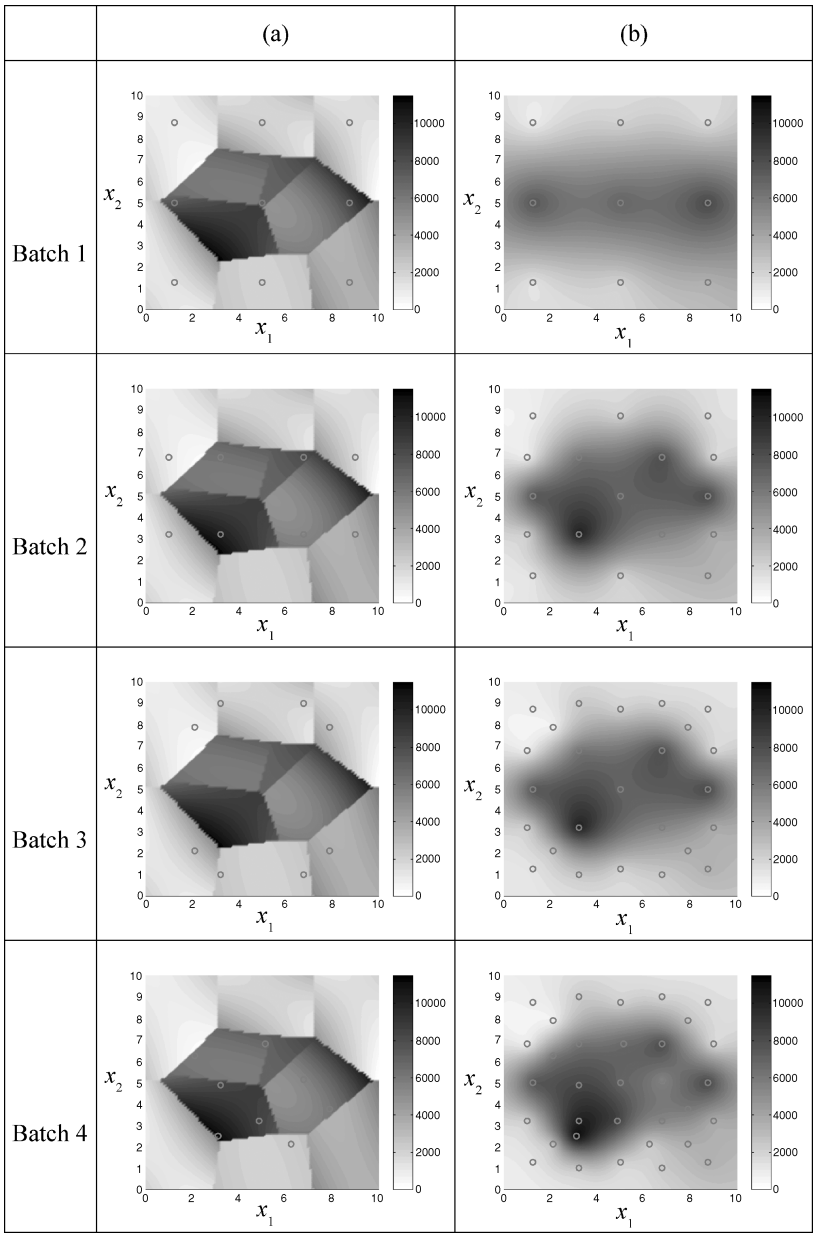

Figure 6. IS process for the 2D-RPV field: (a) contour map of the 2D-RPV field and samples taken at the current batch; (b) contour map of the current model based on the accumulation samples up to the current batch.

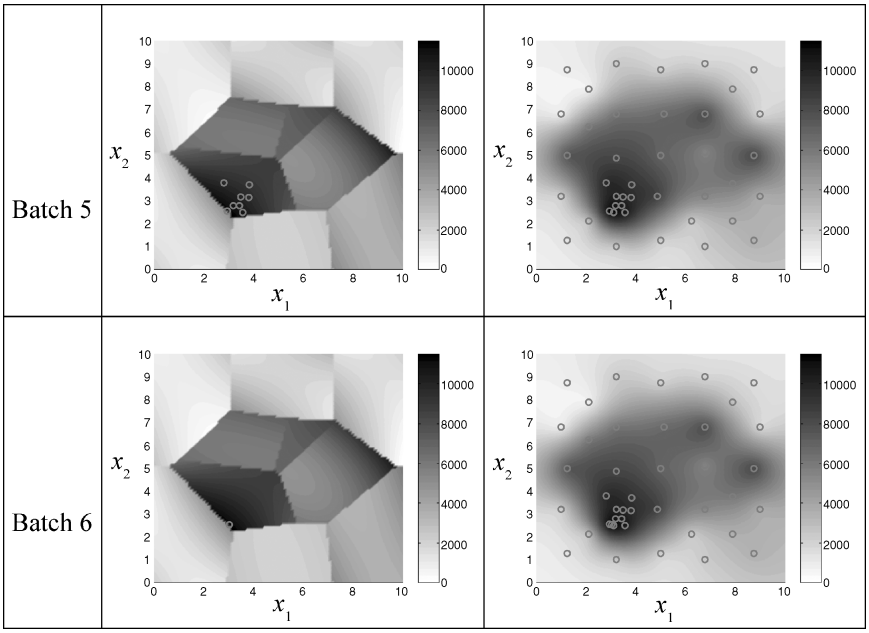

Figure 6. Continued.

but none of the rest RG methods does explains the by-chance phenomena of the RG scheme.

Observing the MAEs of the the whole range (0–100%) as listed in Table 2, we know that the mapping accuracy of both sampling schemes is comparable. While RG of 36, 49 and 64 samples is more accurate in the lower ranges (0–50%), IS of 33, 41 and 42 samples at the end of Batches 4, 5 and 6 is better than RG in the higher ranges(50–100%). It is expected that the overall modeling performance after 33 sample points taken at the end of Batch 4 by IS should be between that of RG-25 and RG-36. When looking closely at the performance of the whole range in Table 2, the IS method is worse than the RG method. The 50–75% range and 75–100% range only occupy 25% of the total population, the accurate modeling performance in these two region are dwarfed by the rough modeling performance in the lower ranges if the overall modeling performance is concerned. The objective of the study is to locate the hot spot regions with the minimum sampling number, the better prediction for the high attribute value area is valuable at the expense of a little rough prediction for the uninterested area. Comparing with Case 1, this case has very rugged landscape with 10 discontinuous zones. The IS method work as expected in both of these two cases even though zero nugget parameter is presumed.

CONCLUSIONS

Hot spots (maxima or extremes) have great significance for mapping the distribution of attributes in an area to be surveyed and for making decisions about the area such as usage of a land, remedy to a polluted district, mining design, etc. An interactive sampling strategy has been proposed in this study for the purpose of locating the hot spot regions of an attribute in a given under-survey area. In the proposed sampling strategy, the ordinary kriging estimator (OK-estimator) is used as a meta-modeling tool to build a response surface of the interested attribute with spatial coordinates from existing sample data. Information analysis is then used to search the response surface for the future sampling locations in order to obtain the rough profile of low attribute value regions and the detailed profile of high attribute value regions.

Two case studies have demonstrated effectiveness of the proposed interactive sampling strategy. Modeling on the 55 field data taken from a heavy metal contamination site generates the first case. The landscape in this case is smooth. The second case with very rugged landscape is a RPV model. For the purpose of comparison, the one-time regular square-grid sampling scheme is used in parallel with the proposed strategy. The results of the two cases show that the proposed IS scheme is able to locate hot spot regions with much less samples in a deterministic way in contrast to the regular square- grid scheme which find hot spot regions by interpolating among a sufficiently large set of samples if we do not consider the good luck of a grid point dropping right into a hot spot region. The simulation results also show that the quality of mapping the whole area of survey are comparable for both IS and RG, though less samples are taken by the proposed IS scheme than the typical RG method. This can be explained by the interpolation nature of the OK-estimator and by the fact that hot spot regions are of significance to the accuracy of kriging mapping.

Though the proposed IS method is effective and efficient in locating hot spot regions, its application is limited to that iterative sampling process is feasible or that chemical analysis is expensive. Some cases in environmental investigation and in geological exploration are suitable for the use of the proposed interactive sampling strategy. Some issues, such as the determination of sample numbers in every batch during iteration, the choice of a meta-modeling tool are as yet to be explored.

ACKNOWLEDGMENTS

The authors thank the computer work provided by Mr. Chi-Shen Tai, and financial support provided by National Science Council, Taiwan, for this work through the grant NSC90-2622-E007-003.

REFERENCES

Brus, D. J., and de Gruijter, J. J., 1977, Random sampling or geostatistical modeling? Choosing between design-based and model-based sampling strategies for soil (with discussion): Geoderma, v. 80, p. 1–59.

Chen, J., Wong, D. S. H., Jang, S.-S., and Yang, S.-L., 1998, Product and process development using artificial neural-network model and information analysis: AIChE J., v. 44, p. 876–887.

Chen, J., Chu, P. P.-T., Wong, D. S. H., and Jang, S.-S., 1999, Optimal design using neural network and information analysis in plasma etching: J. Vac. Sci. Technol. B, v. 17, no. 1, p. 145–153.

Chu, J.-Z., Shieh, S.-S., Jang, S.-S., Chien, C.-I., Wan, H.-P., and Ko, H.-H., 2003, Constrained optimization of combustion in a simulated coal-fired boiler using artificial neural network model and information analysis: Fuel, v. 82, p. 693–703.

Domburg, P., de Gruijter, J. J., and van Beek, P., 1997, Designing efficient soil survey schemes with a knowledge-based system using dynamic programming: Geoderma, v. 75, p. 183–201.

Falcioni, M., and Deem, M. W., 2000, Library design in combinatorial chemistry by Monte Carlo methods: Phys. Rev. E, v. 61, p. 5948–5952.

Haykin, S., 1999, Neural networks: A comprehensive foundation, 2nd edn.: Prentice Hall International, Inc., Englewood Cliffs, NJ.

Isaaks, E. H., and Srivastava, R. M., 1989, Applied geostatistics: Oxford University Press, New York.

Jones, D. R., Schonlau, M., and Welch, W. J., 1998, Efficient global optimization of expensive black-box functions: J. Global Optimization, v. 13, no. 4, p. 455–492.

Juang, K.-W., Lee, D.-Y., and Chen, Z.-S., 1996, Prediction of spatial distribution of heavy metals in contaminated soils by geostatistics: II. Effect of sampling design: Journal. Chinese Agric. Chem. Soc., v. 34, p. 683–694.

Kirkpatrick, S., Gelatt, C. D., Jr., and Vecchi, M. P., 1983, Optimization by simulated annealing: Science, v. 220, p. 671–680.

Lin, J.-S., and Jang, S.-S., 1998, Nonlinear dynamic artificial neural network modeling using an information theory based experimental design approach: Ind. Eng. Chem. Res., v. 37, p. 3640–3651.

Mantoglou, A., and Wilson, J. L., 1982, The turning bands method for simulation of random fields
A1 using line generation by a spectral method: Water Resour. Res., v. 18, p. 1379–1394.

Sasena, M. J., Papalambros, P. Y., and Goovaerts, P., 2000, Metamodeling sampling criteria in a global optimization framework, *in* Proceedings of the 8th AIAA/NASA/USAF/ISSMO Symposium on Multidisciplinary Analysis and Optimization, Paper No. AIAA-2000-4921, Long Beach, CA.

Shannon, C. E., 1948, A mathematical theory of communication: Bell Syst. Tech. J., v. 27, p. 379.

Shannon, C. E., and Weave, W., 1949, The mathematical theory of communication: University of Illinois Press, Urbana, 125 p.

van Groenigen, J. W., 2000, The influence of variogram parameters on optimal sampling schemes for mapping by kriging: Geoderma, v. 97, p. 223–236.

van Groenigen, J. W., Pieters, G., and Stein, A., 2000, Optimizing spatial sampling for multivariate contamination in urban areas: Environmetrics, v. 11, p. 227–244.

van Groenigen, J. W., Siderius, W., and Stein, A., 1999, Constrained optimization of soil sampling for minimization of the kriging variance: Geoderma, v. 87, p. 239–259.

van Groenigen, J. W., Stein, A., and Zuurbier, R., 1997, Optimization of environmental sampling using interactive GIS: Soil Technol., v. 10, p. 83–97.

Watson, A. G., and Barnes, R. J., 1995, Infill sampling criteria to locate extremes: Math. Geol., v. 27, no. 5, p. 589–608.

Yen, C.-H., Wong, D. S. H., and Jang, S.-S., 2003, Information directed sampling for combinatorial material synthesis and library design: J. Chem. Eng. Jpn., v. 36, no. 9, p. 1034–1044.

Queries

A1: This reference is not cited in text. Please check.