From the Thirteenth Meeting of the IFIP Working Group on Reliability and Optimization of Structural Systems, 2006. Kobe, Japan.

Sampling strategies to detect threshold excursions in random fields

Jack W. Baker
Stanford University

Michael H. Faber
Swiss Federal Institute of Technology (ETH Zurich)

ABSTRACT: Sampling strategies are considered for efficiently detecting threshold excursions within realizations of random fields. By considering costs of sampling and costs of non-detection of excursions, a risk-based sampling framework is described. Computation of excursion probabilities, conditional upon observations from samples, is performed by using a conditional simulation technique to generate realizations of random fields that are consistent with the observed values. Simple numerical examples are used to demonstrate the use of the approach, and to illustrate how it might be used to identify adaptive sampling schemes that will aid in efficiently detecting threshold excursions. The problem is motivated by geotechnical sampling problems, where a limited testing budget is available to obtain information about the potential presence of interesting features underground (e.g., pockets of weak or liquefiable soil).

1 INTRODUCTION

The goal of many geotechnical testing activities is to understand the range of potential values that various soil properties might take. This includes intrinsic soil properties such as strength or soil type, as well as levels of contamination in the soil. Because these properties are unknown, they can be represented by random variables. As the properties also tend to be spatially dependent (due to, for example, similarities in materials and deposition patterns), the spatial distribution of soil property values might be represented by a random field.

Common goals in a geotechnical testing activity are to characterize the probability distribution of some soil property (or properties) at a given location, or to estimate the spatial dependence of the soil property [1]. A third goal might be to detect whether the soil property exceeds some given value within the region of interest. This may arise if an engineer is trying to identify areas within a site having low bearing capacity or high liquefaction susceptibility.

In this paper, sampling strategies to most effectively achieve this third goal of detecting excursions are considered. This goal is closely related to search theory, a topic studied in great detail (see, e.g., the survey of [2]). This study varies slightly from most search theory applications, which are concerned with detecting a discrete feature, in that here we are considering extreme values of a continuous random field, so that even non-excursions provide information about the likelihood of an excursion in nearby areas. The fact that all observations provide differing levels of information, however, can greatly increase the numerical complexity of the problem.

Here, a series of simple calculations are performed to illustrate the problem and to point to effective strategies for sampling. With these basic results obtained, potential extensions to more complex situations are discussed.

2 RANDOM FIELD DEFINITION

It is assumed that the property of interest is a Gaussian random field with known mean, variance and autocorrelation. In the example below, the following model is assumed:

$$\mu_F(x,y) = 0 \tag{1}$$

$$\sigma_F(x,y) = 1 \tag{2}$$

$$\rho(\Delta h) = e^{-\Delta h / d} \tag{3}$$

where $\mu_F(x,y)$ and $\sigma_F(x,y)$ are the mean and standard deviation, respectively, of the random field F at location $\{x,y\}$. The term $\rho(\Delta h)$ is the autocorrelation of the random field at two points separated by a distance Δh, and d is a constant controlling the scale over which correlations are significant. The field is assumed to be stationary and isotropic to simplify the presentation, although those conditions can be relaxed without difficulty. The objective is to determine whether a given realization of the random field exceeds a given threshold within some finite domain, using as few samples as possible.

3 RISK-BASED SAMPLING

Although the computational expense can be severe, risk-based sampling approaches provide a useful and rational approach for optimizing sampling strategies [3, 4]. The concept of this approach is to take samples in a manner that minimizes the risk associated with the system. In this way, the cost of sampling and expected cost of potential failures can be minimized. In the most general sampling approach here, the number of samples and locations can be left unspecified, and further samples can be taken as long as the marginal cost of additional samples is less than the marginal value of information from the additional samples. A simpler problem is one where the number of samples is specified, and only the locations of the samples must be chosen. Sampling strategies will also change depending upon whether the sampling is adaptive (i.e., observed results from previous samples can be used when choosing future sample locations) or whether all sample locations must be chosen before sampling begins.

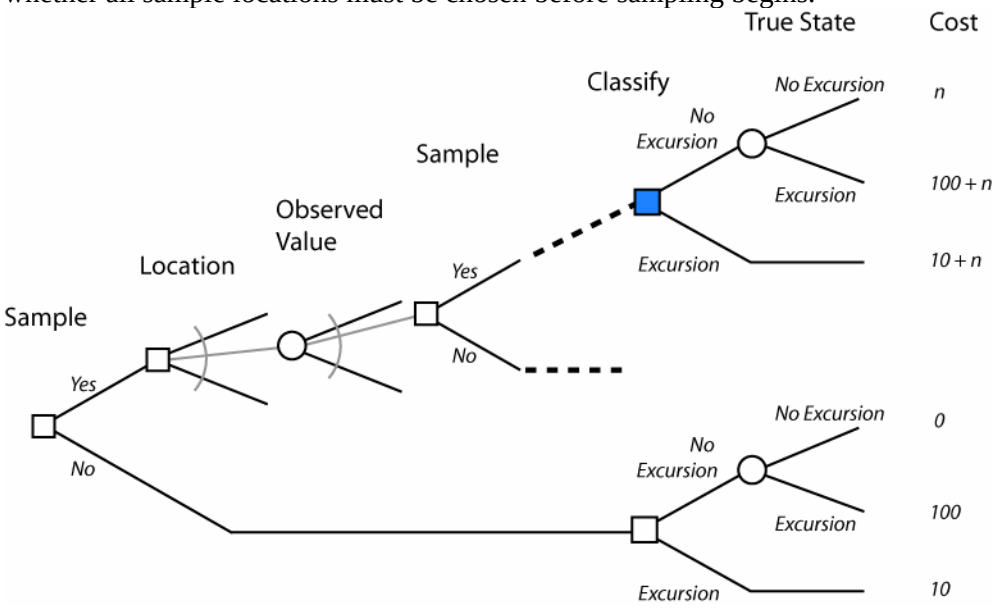

Figure 1: A decision tree for deciding the optimal number of samples and their locations.

To illustrate this sampling approach, we first consider the case where samples are taken until the value of information provided by the sample is less than the cost of the sample. To illustrate, consider an example where each sample has a cost of 1 unit. Classifying a realization as contain-

ing an excursion leads to an additional cost of 10 units, due to mitigation measures that are needed; this cost is incurred regardless of whether an excursion actually exists. Failing to detect an excursion has a cost of 100, which represents the cost of a delay during construction to take remedial actions, or the cost of a failure in the future.

A decision tree for this example is shown in Figure 1. The classification decision (the shaded decision node in Figure 1) is straightforward to calculate. Because the number of samples and their locations is specified, it only remains to compute the probability of an excursion, denoted p, conditional upon the observed sample values (a method for computing this probability is discussed in the following section). For this example, if $p<0.1$ then the minimum expected cost is obtained by classifying the site as having no excursions ($E[cost] = 100*p + n$). If $p>0.1$ then the minimum expected cost is obtained by classifying the site as having an excursion ($E[cost] = 10 + n$). For $p=0.1$, the two choices are equivalent.

Pre-posterior decision analysis can in principle be used to determine the number of samples to be taken and their locations [5]. In practice, however, computing the costs required to evaluate this decision is very difficult. The value of information obtained from an additional sample must be maximized by optimally choosing the location at which the sample should be taken. But the location at which the sample should be taken, and its value, depends upon the number of further samples that might be taken (i.e., if this is the last sample, it should be placed in the middle of an area of interest, but if two more samples are to be taken, then they might be spaced to provide more uniform coverage).

Finding the optimal sampling strategy requires evaluating every possible path in Figure 1, because decision path will depend upon the (random) observed values of future samples. As a consequence of this, the problem differs from deterministic dynamic programming problems, where many branches do not need to be evaluated in order to locate the optimal decision path. Using symmetry arguments, some of the potential combinations are equivalent and the number of evaluations can be reduced. But because the excursion probabilities needed for formulating decision rules are calculated based on numerical simulations, it is prohibitively computationally expensive to evaluate decision trees having a large number of possible paths.

To simplify the decision problem, one could first fix the allowed number of samples so that the decision problem only involves choosing the locations of these samples. This is perhaps a realistic situation for many practical problems (e.g., the budget for testing is fixed before testing begins, or the sampling equipment is mobilized for a fixed amount of time). Because the number of possible decisions is greatly reduced in this case, it is possible to perform some calculations.

3.1 *Conditional excursion probability, given observations*

When deciding whether to classify a site as having an excursion or not, it is necessary to compute the probability of a threshold excursion, given the observed values from completed tests. Although many analytical results have been derived relating to properties of random fields [6-8], no methods exist for computing excursion probabilities conditional upon observations. Thus a conditional simulation procedure is used here. This procedure can be used to simulate realizations of the random field that are consistent with the specified mean, variance and autocorrelation of the field, but that also agree perfectly with observed values at specified locations (see, e.g., [9]). The procedure takes advantage of the fact that conditional distributions of joint standard Gaussian random variables are also Gaussian with the following distribution

$$\left(\mathbf{Y} \mid \mathbf{Y}_{obs} = \mathbf{y}\right) \sim N\left(\boldsymbol{\Sigma}_{12} \cdot \boldsymbol{\Sigma}_{22}^{-1} \cdot \mathbf{y},\ \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12} \cdot \boldsymbol{\Sigma}_{22}^{-1} \cdot \boldsymbol{\Sigma}_{21}\right) \tag{4}$$

where $\mathbf{y}$ is the vector of observed values from previous measurements, $\mathbf{Y}$ is the vector of random variables at unobserved locations in the area of interest, $\boldsymbol{\Sigma}_{11}$ is the covariance matrix of the unobserved values $\mathbf{Y}$, $\boldsymbol{\Sigma}_{22}$ is the covariance matrix of the random variables associated with the observed locations, and $\boldsymbol{\Sigma}_{12}$ and $\boldsymbol{\Sigma}_{21}$ are the cross-covariance matrices between the random field values at the observed and unobserved locations. This approach is explained in more detail elsewhere [9].

It is in general not possible to compute excursion probabilities directly from the result of equation (4). One can, however, simulate samples from this distribution, and then count the number of samples that contain an excursion. Using this Monte Carlo approach, it is possible to

estimate excursion probabilities conditional upon any set of observed values at any locations of interest. Example conditional simulations in one dimension are shown in Figure 2b.

4 EXAMPLE CALCULATIONS

Several simple examples are described in the following sections, to illustrate how the above framework can be applied.

4.1 *One-dimensional sample space*

The simplest possible example is to first consider only a one-dimensional domain of interest, and to consider a small number of samples in that area. The region of interest is assumed to be the interval from 0 to 100, and the random function is defined by Equations 1 through 3, with the distance parameter *d* equal to 25. The threshold level of interest is set equal to 2. The cost of classifying a site as having an excursion was set equal to 40, while the cost of not identifying an excursion is equal to 100 and the cost of samples was neglected because the number of samples is constant. An example realization is shown in Figure 2a, and conditional simulations of this one-dimensional random process are shown in Figure 2b. Because the marginal distribution of individual locations is standard Gaussian, their individual exceedance probabilities are equal to $1-\Phi(2)=0.023$, where $\Phi(\cdot)$ is the standard Gaussian cumulative distribution function. The probability that a given realization of the field will contain an excursion is estimated using Monte Carlo simulation to be 0.304.

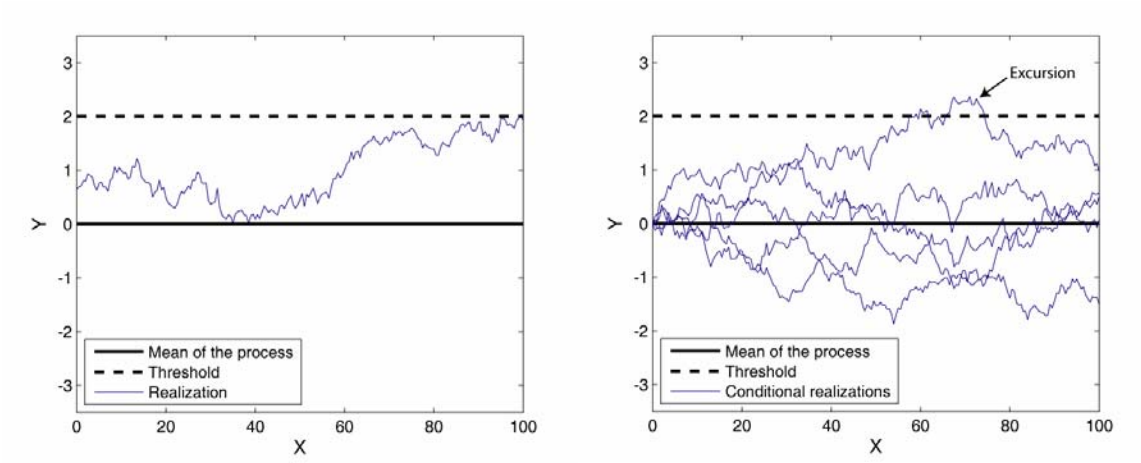

Figure 2: (a) A single realization of the random process in one dimension. (b) Five realizations of the random process, conditional upon observing a value of 0 at location X = 0.

This example can be used to confirm the potential benefit of an adaptive sampling scheme. Figure 3 shows expected costs, given that a second sample is taken after observing the value at location X=0. The costs are conditional upon the observed value at the first location. If the first location is observed to have a value much below the threshold (i.e., on the left side of Figure 3), then the cost is minimized by taking the second sample at a location far from the location of the first sample. If, however, the first sample yields an observation near to the threshold value, then the cost is minimized by taking an additional sample near to the original sample. This makes intuitive sense: if the first sample suggests that the surrounding region is "safe" (far below the threshold) then additional samples should be taken elsewhere, but if the first sample indicates a potential nearby excursion, then additional samples at nearby locations will be useful.

By accounting for the probability distribution of the observed value at the first sample location, and assuming that the second sample is taken at whichever of the three locations minimizes the expected cost, one can compute that the expected cost of this adaptive sampling scheme is 27.1. In contrast taking two samples at {0, 10}, {0, 40}, and {0, 70} yields expected costs of 29.3, 27.5, and 28.0, respectively. Thus, as would be expected intuitively, the adaptive sampling scheme yields lower expected costs than the alternative non-adaptive schemes. The relative benefit of adaptive sampling schemes is expected to increase significantly as the number of samples increases.

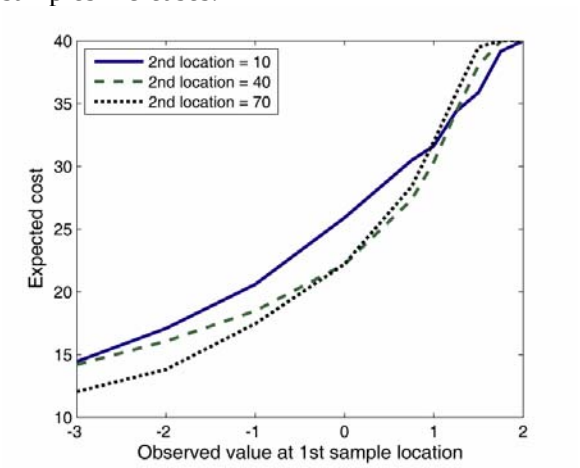

Figure 3: Expected cost of a second sample, conditional upon the observed value at the first location. The first sample is taken at X=0, and results are shown for three potential second locations.

4.2 *Two-dimensional sample space*

In this example, the sample region of interest is a surface of dimension 100 x 100, composed of discrete cells each having dimension 1 x 1. Examples of the simulations are shown in Figure 4. The Gaussian field was defined according to equations 1 through 3, and a threshold level of 3 was chosen. The marginal threshold exceedance probabilities of individual locations are equal to 0.0013, and the probability that a given realization of the field will contain an excursion is 0.28.

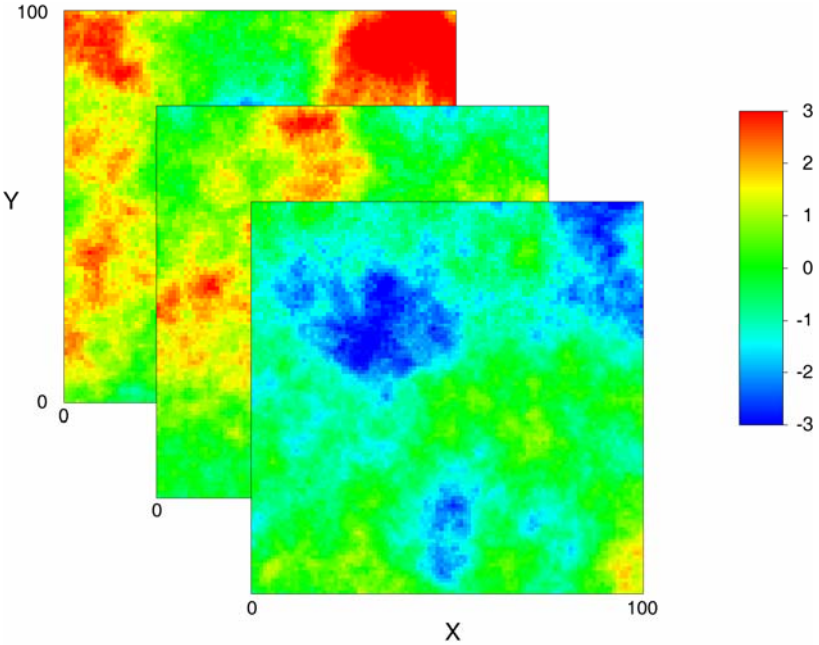

Figure 4: Example simulations of the random field used for calculations.

Direct simulation of these realizations is possible, but requires inversion of a 10,000 x 10,000 covariance matrix. To save computational expense, a conditional simulation approach was used (see [10] for details). Figure 5 shows the excursion regions of the simulations from Figure 4.

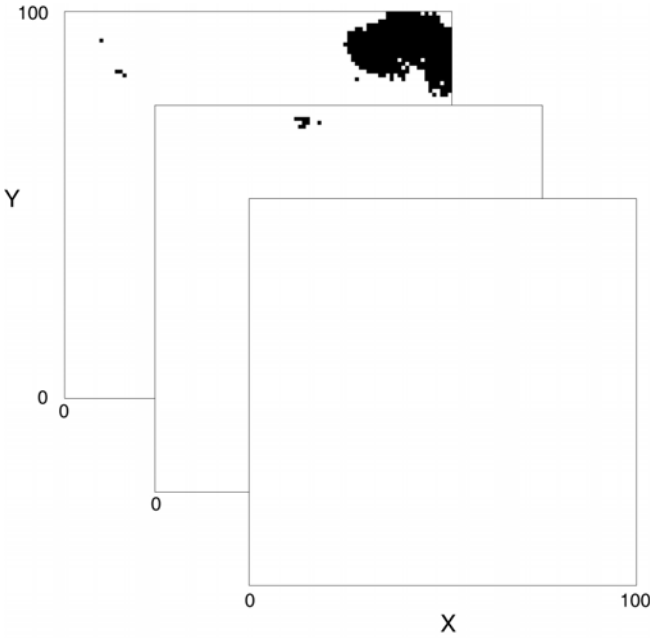

Figure 5: Threshold excursions of the random field simulations shown above.

4.2.1 *One sample*

When only one sample is to be taken in the region, intuition suggests that it should be taken in the middle of the site. To confirm this using the proposed procedure, six sample locations are considered: in each case the *y* value of the sample location is 50, and *x* values of 0, 10, 20, 30, 40 and 50 are considered, as illustrated in Figure 6a. Because of symmetry, *x* values greater than 50 are not needed, and the *x* and *y* coordinates can be interchanged to reflect sampling along the vertical axis.

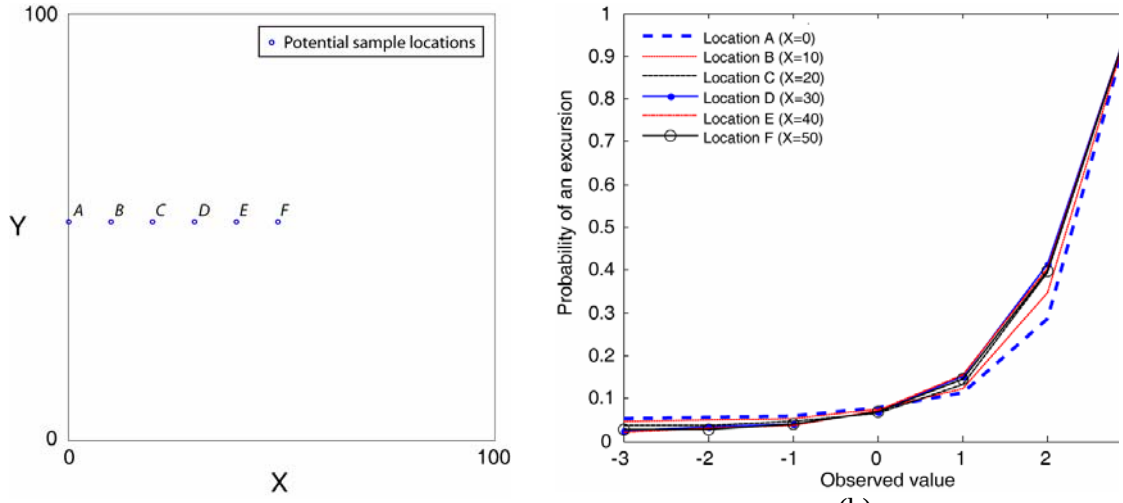

Figure 6: (a) Illustration of the six potential sample locations to be considered. (b) Probability of an excursion, conditional upon the observed value at the potential sample location.

For each sample location and potential observed value, simulations of conditional realizations can be used to compute the probability of an excursion. The results for a variety of observed values and sample locations are shown in Figure 6b. By using the excursion probabilities from Figure 6b, it is possible to compute the conditional expected cost conditional upon observing the specified value. Then, integrating all possible observed values and weighting each by its probability of occurrence, unconditional expected costs can be computed. The expected costs associated with taking samples at locations A through F are: 8.84 8.50 8.04 7.99 7.93 and 7.92. This indicates that the optimal strategy would be to take a sample near the center of the area of interest, as would be expected.

4.2.2 *Two samples*

If two samples are to be taken, an adaptive procedure can be used to choose the second sample location based on the result of the first location. To keep the number of considered situations manageable, the following system is proposed. First a sample is taken at location $\{x,y\} = \{25,25\}$. In the adaptive sampling situation, the second location can be chosen once the outcome of the first sample is known. If the first sample is close to the excursion threshold, the second sample might be taken near to the first sample in order to learn more about this potentially critical region. If the first sample is far below the threshold, then the second sample should be taken at a location far from the first sample, under the assumption that the region near the first sample is of little concern, so other regions of the site should be studied. The considered locations for the second sample are $\{x,y\} = \{30,30\}$, $\{60,60\}$ and $\{90,90\}$.

Conditional excursion probabilities are computed using the conditional simulation technique described above, and example results are shown in Figure 7 for a variety of sample locations and observed values. These can again be linked with expected costs associated with observing these values, and then by integrating over the probability of observing these values, expected costs of various sample strategies can be evaluated.

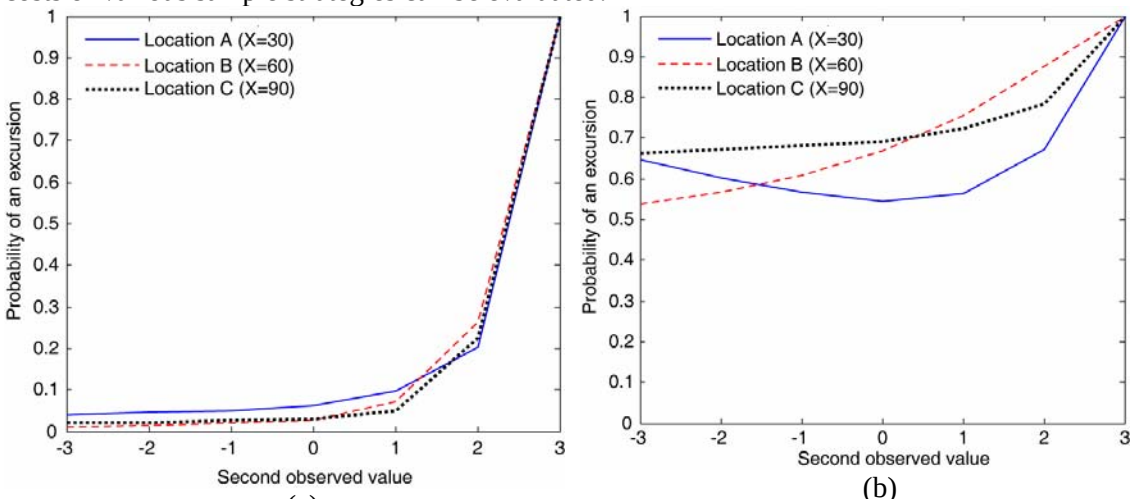

Figure 7: Probability of an excursion, conditional upon the observed values at the potential sample locations. (a) Results given that a value of -2.5 was observed at the first sample location. (b) Results given that a value of 2.5 was observed at the first sample location.

5 FUTURE WORK

The above work illustrates an approach by which decision theory can be used to identify optimal schemes for sampling to detect an excursion in a random field. More work is required, however, before the approach is ready for practical applications with large numbers of samples. Evaluation of the decision tree used to choose sampling strategies requires computation of excursion probabilities conditional upon every potential permutation of sample locations, and this is computationally prohibitive because the probabilities are computed using simulation techniques. An approach will be developed to sample a representative subset of possible sample locations, in

order to identify near-optimal strategies at a fraction of the computational cost. Analytic approximations to excursion probabilities will also be used to efficiently process large numbers of potential sample locations, and to identify the sample options where more careful (i.e., more computationally expensive) evaluations are needed.

The above approach focused only on detecting threshold excursions, while assuming that the properties of the random field were known. In reality, however, many sampling activities are concerned both with quantifying the relevant parameters of the random field as well as detecting excursions. Work relating to sampling strategies for characterizing random fields has been done (e.g., [11]), and those concepts could be combined with the ideas presented here. Evaluating tradeoffs between multiple objectives may be difficult, however.

6 CONCLUSIONS

A risk-based sampling framework for detecting excursions in realizations of random fields has been discussed. The goal of the framework is to identify optimal strategies for choosing sample locations to minimize expected costs, where there are costs associated with taking samples, with misclassifying a site as having no excursions and with classifying a site as having an excursion. The problem is motivated by, for example, geotechnical sampling problems, where a limited testing budget is available to obtain information about the potential presence of interesting features underground (e.g., pockets of weak or liquefiable soil).

Using decision theory, a decision tree was developed to frame this question. A critical need for evaluating the decision tree is an estimate for the probability of an excursion existing in a given sample realization, conditional upon the observed values from samples taken at a specified set of locations. This excursion probability was computed using a conditional simulation approach that allows for a great deal of flexibility in terms of the definition of the random field, the sample locations, and the shape and size of the region of interest. The approach should prove especially useful for developing adaptive sampling strategies, where the location of additional samples is dependent upon the observed values of previous samples. Simple numerical examples were presented to illustrate how this approach could be evaluated. By providing a quantitative method for evaluating these strategies, cost-saving improvements over typical sampling strategies can be formulated.

REFERENCES

[1] Baecher GB, Christian JT. Reliability and statistics in geotechnical engineering. Chichester, West Sussex, England; Hoboken, NJ: J. Wiley; 2003.

[2] Benkoski SJ, Monticino MG, Weisinger JR. A survey of the search theory literature. Naval Research Logistics 1991; 38469-94.

[3] Straub D., Faber M.H. Risk based inspection planning for structural systems. Structural Safety 2005; 27(4):335-55.

[4] Faber M.H., Straub D., Maes M.A. A computational framework for risk assessment of RC structures using indicators. Computer-Aided Civil and Infrastructure Engineering 2006; 21(3):216-30.

[5] Benjamin JR, Cornell CA. Probability, statistics, and decision for civil engineers. New York: McGraw-Hill; 1970.

[6] Vanmarcke E. Random fields, analysis and synthesis. Cambridge, Mass.: MIT Press; 1983.

[7] Adler RJ. The geometry of random fields. Chichester [Eng.] ; New York: J. Wiley; 1981.

[8] Adler RJ, Taylor JE. Random fields and geometry. (in press, preprint at http://iew3.technion.ac.il/~radler/publications.html); 2006.

[9] Baker JW, Faber M. Accounting for soil spatial variability when assessing liquefaction risk Journal of geotechnical and geoenvironmental engineering 2006; (in review).

[10] Deutsch CV, Journel AG. GSLIB geostatistical software library and user's guide. Version 2.0. ed. New York: Oxford University Press; 1997.

[11] Degroot DJ, Baecher GB. Estimating autocovariance of in-situ soil properties. Journal of Geotechnical Engineering 1993; 119(GT1):147-66.