The Alliance of Laboratories in Europe for Research and Technology

ALERT Doctoral School 2014

Stochastic Analysis and Inverse Modelling

Editors:

Michael A. Hicks

Cristina Jommi

Editorial

The ALERT Doctoral School (European Graduate School) in 2014 is organized by Michael A. Hicks and Cristina Jommi from the TU Delft. With the topic *Stochastic Analysis and Inverse Modelling* they have chosen a subject which is without any doubts important but at the same time somehow feared. We know that the ground properties and behaviour are hardly deterministic and in many cases we miss important parameters which can be obtained only by inverse analysis. Still, only in relatively rare occasions we involve appropriate stochastic and inverse techniques. I hope that this school book can change the attitude of the doctoral students attending the school and of those who will read the book afterwards. As usual, the pdf file of the book can be downloaded for free from the website of ALERT Geomaterials – `http://alertgeomaterials.eu`.

On behalf of the ALERT Board of Directors and of all the members of ALERT, I wish you a successful ALERT Doctoral School 2014. I highly appreciate the commitment of the school organizers and contributors of this printed volume and wish them a fruitful event in the Paul Langevin Centre in Aussois.

Ivo Herle
Director of ALERT Geomaterials
Technische Universität Dresden

Contents

Stochastic Analysis and Inverse Modelling: Foreword

This book contains a collection of notes to accompany the lectures of the 2014 ALERT Geomaterials Doctoral School on "Stochastic Analysis and Inverse Modelling". The School has been organized by Michael Hicks and Cristina Jommi (Delft University of Technology), who gratefully acknowledge the hard work of all contributors. The book contains 11 chapters.

Chapter 1 has been written by Gordon Fenton, and contains a review of the basics of probability theory. This includes the theory of random variables, their main characteristics, the effects of dependencies between random variables, and the most common discrete and continuous distributions. Chapters 2-4 have been written by Abed Soubra and Emilio Bastidas. They cover functions of random variables, as well as reliability analysis methods for providing a framework to account for uncertainties in engineering design. These include the First Order and Second Order Reliability Methods, as well as more advanced methods. There then follow a series of chapters focussing on matters relating to the spatial variation of geotechnical properties. Chapters 5-7 have been written by Gordon Fenton and cover random fields, best linear unbiased estimation, and methods of simulating ground property random fields. These are followed by Chapter 8, written by Michael Hicks, focussing on the application of the Random Finite Element Method for assessing the influence of soil heterogeneity on soil behaviour and geotechnical performance. The final 3 chapters are devoted to inverse modelling. Chapter 9 has been written by Alberto Ledesma and describes the use of a Maximum Likelihood approach for back-analysing geotechnical model parameters from field measurements. Chapter 10 has been written by Michele Calvello and includes the use of inverse modelling for parameter estimation, physical modelling, and field monitoring and construction. Finally, Chapter 11 introduces sequential data assimilation and filtering for time-dependent problems and has been written by Cristina Jommi and Patrick Arnold.

We hope that students and researchers will find this collection of notes a useful source of information, both in complementing the lectures of the Doctoral School and as a future source of reference.

Michael HICKS
Cristina JOMMI

Review of Probability Theory

Gordon A. Fenton

Dalhousie University, Canada

Regulatory bodies are increasingly asking geotechnical engineers to provide rational risk assessments to accompany their designs. In order to provide these assessments, geotechnical engineers need a good understanding of both basic probability theory and the more sophisticated, but realistic, random field soil models. This chapter lays the groundwork for this understanding. Starting with the basics of probability, the reader is lead through the theory of random variables, their main characteristics, the effect of dependencies between random variables, and finally reviews the most common discrete and continuous distributions, including extreme value distributions.

1 Basic Probability Concepts

Games of chance played an important role in the development of probability theory. The great gambling houses of Europe frequently hired mathematicians to help improve their own odds over the last 600 years. The mathematical theory of probability was started by Pascal and Fermat, who were French mathematicians of the 1600's, spurred in large part by financing from the gambling house owners.

Although the term *probability* is commonplace today, its exact definition is still a highly controversial subject. There are three 'accepted' interpretations;

1. *equilikely interpretation* - that all outcomes are equally likely
2. *frequency interpretation* - that probability is proportional to the frequency of occurrence,
3. *subjective interpretation* - that probability is derived from experience

The equilikely interpretation is the simplest, although some people argue that randomness can be characterized by the equilikely interpretation only if the 'correct' set of outcomes can be defined. However, if an experiment can result in any one of N different but equally likely outcomes, and if exactly m of these outcomes correspond to the event A, then the probability of event A is m/N under this interpretation.

The frequency interpretation is probably the most powerful and commonly used interpretation of probability. It allows the estimation of probability by counting the number of occurrences of a particular event of interest and dividing by the total number of occurrences possible. For example, if an experiment has two possible outcomes, A and B, and out of 80 experiments, event A occurred 20 times, then we say that the probability of the event A is (at least approximately) 20/80.

Finally, subjective interpretations are often extremely valuable, especially in geotechnical engineering, since years of experience can rarely be captured by a mathematical model... In fact there are those that argue that most engineering probability estimates are subjective. They are probably right in that most engineering situations do not allow a very large number of experiments to assess probabilities according to the frequency interpretation. Often it boils down to an expert's opinion regarding the basic event probabilities, usually derived from years of experience with similar events.

2 Mathematics of Probability

Definition of Probability:

> The probability of an event A, denoted by $\mathrm{P}\big[A\big]$, is a number which satisfies axioms 1 and 2 listed next. More generally, the probabilities associated with any set of disjoint events A_1, A_2, ..., where each $A_i \in S$, are numbers which satisfy axioms 1, 2, and 3.

Three Axioms (fundamental assumptions)

1. for any event A, $\mathrm{P}\big[A\big] \geq 0$
2. a certain event has probability 1: $\mathrm{P}\big[S\big] = 1$
3. for any sequence of disjoint events, A_1, A_2, ...,

$$\mathrm{P}\big[A_1 \cup A_2 \cup \cdots\big] = \mathrm{P}\big[A_1\big] + \mathrm{P}\big[A_2\big] + \ldots \tag{1}$$

From these three fundamental assumptions, all probability theory is constructed!

Important Results:

a) $\mathrm{P}\big[\phi\big] = 0$

b) if A_1 and A_2 are *disjoint* then

$$\mathrm{P}\big[A_1 \cup A_2\big] = \mathrm{P}\big[A_1\big] + \mathrm{P}\big[A_2\big] \qquad \text{(addition rule)} \tag{2}$$

and similarly for any finite number of disjoint events.

c) $\mathrm{P}\big[A \cup A^c\big] = \mathrm{P}\big[A\big] + \mathrm{P}\big[A^c\big]$ since A and A^c are disjoint events. Furthermore

$$\mathrm{P}\big[A \cup A^c\big] = \mathrm{P}\big[S\big] = 1 \quad \rightarrow \quad \mathrm{P}\big[A^c\big] = 1 - \mathrm{P}\big[A\big] \tag{3}$$

d) $0 \leq \mathrm{P}\big[a\big] \leq 1$

e) if $a \subset b$ then $\mathrm{P}\big[a\big] \leq \mathrm{P}\big[b\big]$

f) for any two events a and b,

$$\mathrm{P}\big[A \cup B\big] = \mathrm{P}\big[A\big] + \mathrm{P}\big[B\big] - \mathrm{P}\big[A \cap B\big] \tag{4}$$

g) similarly for any three events A, B, and C,

$$\begin{aligned}\mathrm{P}\big[A \cup B \cup C\big] = \mathrm{P}\big[A\big] + \mathrm{P}\big[B\big] + \mathrm{P}\big[C\big] - \mathrm{P}\big[A \cap B\big] \\ - \mathrm{P}\big[A \cap C\big] - \mathrm{P}\big[B \cap C\big] + \mathrm{P}\big[A \cap B \cap C\big]\end{aligned} \tag{5}$$

2.1 Conditional Probability

The probability of an event is often affected by the occurrence of other events and/or the knowledge of information relevant to the event. Given two events A_1 and A_2 resulting from an experiment, the conditional probability of A_1 occurring *given* that we know A_2 has occurred is denoted as $\mathrm{P}\big[A_1 \,|\, A_2\big]$. Note that the vertical bar, $|$, means 'given that'.

The conditional probability can be interpreted graphically using a Venn diagram. If we know A_2 has occurred, then the outcome must lie somewhere in A_2. That is,

- The 'new' sample space is A_2
- The probability that the outcome lies in $A_1 \cap A_2$ is just the area of $A_1 \cap A_2$ divided by the new sample space area A_2, ie

$$\mathrm{P}\big[A_1 \,|\, A_2\big] = \frac{\mathrm{P}\big[A_1 \cap A_2\big]}{\mathrm{P}\big[A_2\big]} \tag{6}$$

This leads to the **multiplication rule:**

$$\mathrm{P}\big[A_1 \cap A_2\big] = \mathrm{P}\big[A_1 \,|\, A_2\big] \cdot \mathrm{P}\big[A_2\big] \;=\; \mathrm{P}\big[A_2 \,|\, A_1\big] \cdot \mathrm{P}\big[A_1\big] \tag{7}$$

Note that the addition rule can be applied conditionally, ie

$$\mathrm{P}\big[A_1 \cup A_2 \,|\, E\big] = \mathrm{P}\big[A_1 \,|\, E\big] + \mathrm{P}\big[A_2 \,|\, E\big] - \mathrm{P}\big[A_1 \cap A_2 \,|\, E\big] \tag{8}$$

where $\mathrm{P}\big[A_1 \cap A_2 \,|\, E\big] = \mathrm{P}\big[A_1 \,|\, A_2 \cap E\big] \cdot \mathrm{P}\big[A_2 \,|\, E\big]$.

More generally, all probabilities are conditional probabilities. It's just that we don't usually express the more obvious ones. For example, if I compute the probability that a stock market price will fall below \$3.50, the resulting probability is implicitly conditional on a variety of things which (usually) have probability 1.0; for example,

the computed estimate will be conditional on the fact that we have developed the concept of money, that a stock market exists, that the world has not been destroyed by a supernova, etc. The point is that any of the probability relationships developed above and below have the same form regardless of the number of conditional events appearing to the right of the | sign. For example, the relationship

$$\mathrm{P}\big[A \cup B\big] = \mathrm{P}\big[A\big] + \mathrm{P}\big[B\big] - \mathrm{P}\big[A \cap B\big] \tag{9}$$

is still valid if we add any number of conditional events, as in

$$\mathrm{P}\big[A \cup B \,|\, C \cap D\big] = \mathrm{P}\big[A \,|\, C \cap D\big] + \mathrm{P}\big[B \,|\, C \cap D\big] - \mathrm{P}\big[A \cap B \,|\, C \cap D\big] \tag{10}$$

That is, every term just has the conditions added to it.

Note that conditional events *do not* obey the rules of probability. Conditional events are assumed to have occurred – they are non-random. For example, it is a *mistake* to assume that $\mathrm{P}\big[A \,|\, B \cup C\big] = \mathrm{P}\big[A \,|\, B\big] + \mathrm{P}\big[A \,|\, C\big] - \mathrm{P}\big[A \,|\, B \cap C\big]$. If you do need to break the event $B \cup C$ up in some way, you need to first turn it around. For example, Bayes' Theorem allows us to write

$$\mathrm{P}\big[A \,|\, B \cup C\big] = \frac{\mathrm{P}\big[B \cup C \,|\, A\big]\,\mathrm{P}\big[A\big]}{\mathrm{P}\big[B \cup C\big]} \tag{11}$$

where **now** we can write

$$\mathrm{P}\big[B \cup C\big] = \mathrm{P}\big[B\big] + \mathrm{P}\big[C\big] - \mathrm{P}\big[B \cap C\big] \tag{12}$$

and

$$\mathrm{P}\big[B \cup C \,|\, A\big] = \mathrm{P}\big[B \,|\, A\big] + \mathrm{P}\big[C \,|\, A\big] - \mathrm{P}\big[B \cap C \,|\, A\big] \tag{13}$$

2.2 Statistical Independence

If the occurrence (or non-occurrence) of one event does not affect the probability of another event, then the two events are called statistically independent (**this is not the same as disjoint!**). For example if event A_1 is independent of event A_2 then

$$\begin{aligned} \mathrm{P}\big[A_1 \,|\, A_2\big] &= \mathrm{P}\big[A_1\big] \\ \mathrm{P}\big[A_2 \,|\, A_1\big] &= \mathrm{P}\big[A_2\big] \end{aligned} \tag{14}$$

so that $\mathrm{P}\big[A_1 \cap A_2\big] = \mathrm{P}\big[A_1 \,|\, A_2\big] \cdot \mathrm{P}\big[A_2\big] \;=\; \mathrm{P}\big[A_1\big] \cdot \mathrm{P}\big[A_2\big]$

Similarly for three mutually statistically independent events we can write

$$\begin{aligned} \mathrm{P}\big[A_1 \cap A_2 \cap A_3\big] &= \mathrm{P}\big[A_1 \,|\, A_2 \cap A_3\big] \cdot \mathrm{P}\big[A_2 \cap A_3\big] \\ &= \mathrm{P}\big[A_1 \,|\, A_2 \cap A_3\big] \cdot \mathrm{P}\big[A_2 \,|\, A_3\big] \cdot \mathrm{P}\big[A_3\big] \\ &= \mathrm{P}\big[A_1\big] \cdot \mathrm{P}\big[A_2\big] \cdot \mathrm{P}\big[A_3\big] \end{aligned} \tag{15}$$

since we must have $\mathrm{P}\big[A_1 \,|\, A_2 \cap A_3\big] = \mathrm{P}\big[A_1\big]$ and $\mathrm{P}\big[A_2 \,|\, A_3\big] = \mathrm{P}\big[A_2\big]$ if A_1, A_2, and A_3 are statistically independent.

Note that A_1, A_2, and A_3 may be pairwise independent and yet not mutually independent (in which case $\mathrm{P}\big[A_1 \,|\, A_2 \cap A_3\big] \neq \mathrm{P}\big[A_1\big]$ even though $\mathrm{P}\big[A_i \,|\, A_j\big] = \mathrm{P}\big[A_i\big]$ for all $i \neq j$).

2.3 Total Probability and Event Trees

Sometimes the probability of an event E cannot be determined directly, its probability being given in terms of the occurrence of other events.

If $A_1, A_2, \ldots, A_n$ form a *partition* of the sample space S (ie. are *mutually exclusive* and *collectively exhaustive*), then we can express the probability of E as a sum of conditional probabilities;

$$\begin{aligned}\mathrm{P}\big[E\big] &= \mathrm{P}\big[E \cap S\big] \;=\; \mathrm{P}\big[E \cap (A_1 \cup A_2 \cup \cdots \cup A_n)\big] \\ &= \mathrm{P}\big[(E \cap A_1) \cup (E \cap A_2) \cup \cdots \cup (E \cap A_n)\big] \\ &= \mathrm{P}\big[E \cap A_1\big] + \mathrm{P}\big[E \cap A_2\big] + \cdots + \mathrm{P}\big[E \cap A_n\big] \qquad (16)\end{aligned}$$

To obtain the last line, we note that because $A_1, A_2, \ldots, A_n$ are disjoint, the events $(E \cap A_1), (E \cap A_2), \ldots, (E \cap A_n)$ must also be disjoint, so that the probability can be written as a simple sum.

Now since $\mathrm{P}\big[E \cap A_i\big] = \mathrm{P}\big[E \,|\, A_i\big] \cdot \mathrm{P}\big[A_i\big]$, this can be written in the form

$$\mathrm{P}\big[E\big] = \mathrm{P}\big[E \,|\, A_1\big]\,\mathrm{P}\big[A_1\big] + \mathrm{P}\big[E \,|\, A_2\big]\,\mathrm{P}\big[A_2\big] + \cdots \mathrm{P}\big[E \,|\, A_n\big]\,\mathrm{P}\big[A_n\big] \qquad (17)$$

which is the Total Probability Theorem.

We can illustrate this Theorem through the following example.

A company manufactures network cards of which 50% are produced at plant A, 30% at plant B, and 20% at plant C. It is known that 1% of plant A's, 2% of plant B's, and 3% of plant C's output are defective. What is the probability that a network card chosen at random will be defective?

Solution:

Let A be the event that the network card was produced at plant A.
Let B be the event that the network card was produced at plant B.
Let C be the event that the network card was produced at plant C.
Let D be the event that the network card is defective.

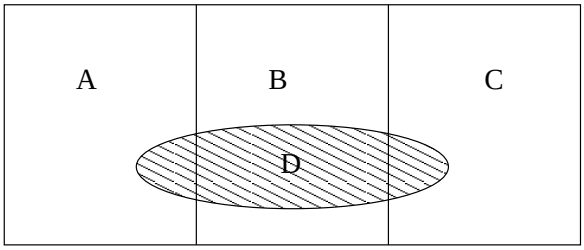
Given:

$$\begin{aligned}\mathrm{P}\big[A\big] &= 0.50, & \mathrm{P}\big[B\big] &= 0.30, & \mathrm{P}\big[C\big] &= 0.20, \\ \mathrm{P}\big[D|A\big] &= 0.01, & \mathrm{P}\big[D|B\big] &= 0.02, & \mathrm{P}\big[D|C\big] &= 0.03.\end{aligned}$$

We are looking for: $\mathrm{P}\big[D\big]$.

Two possible approaches:

Approach 1
A Venn diagram of the sample space appears as follows,



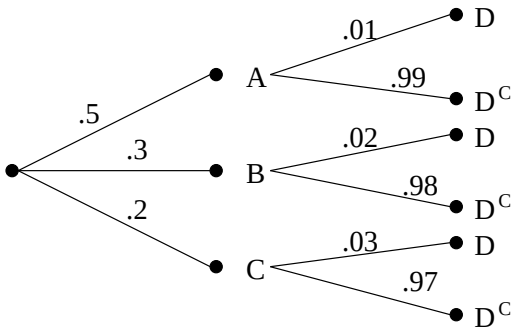
The information given in the problem is not easily portrayed in a Venn diagram. However, the event of interest has been shaded in the diagram, and $\mathrm{P}\big[D\big]$ can be computed as follows,

$$\begin{aligned}
\mathrm{P}\big[D\big] &= \mathrm{P}\big[(A \cap D) \cup (B \cap D) \cup (C \cap D)\big] \\
&= \mathrm{P}\big[A \cap D\big] + \mathrm{P}\big[B \cap D\big] + \mathrm{P}\big[C \cap D\big] \\
&\quad \text{since } A \cap D, B \cap D, \text{ and } C \cap D \text{ are mutually exclusive} \\
&= \mathrm{P}\big[D|A\big]\,\mathrm{P}\big[A\big] + \mathrm{P}\big[D|B\big]\,\mathrm{P}\big[B\big] + \mathrm{P}\big[D|C\big]\,\mathrm{P}\big[C\big] \\
&= 0.01(0.5) + 0.02(0.3) + 0.03(0.2) \\
&= 0.017 \qquad (18)
\end{aligned}$$

Approach 2
When we have only unconditional probabilities like $\mathrm{P}\big[A\big]$, $\mathrm{P}\big[B\big]$, ..., etc., it is relatively easy to draw these probabilities on a Venn diagram. However, since conditional probabilities are ratios of the areas in the diagram, they are less easily seen. Conditional probabilities find a more natural home in event trees. Event trees must be constructed carefully according to certain rules. The basic idea is that there is a *starting node* from which two or more branches leave. At the end of each of these branches there is another node from which more branches may leave (and go to more separate nodes). The idea is repeated from each node as often as required to completely depict all possibilities. For any node other than the starting node, the branches leaving the node will hold *conditional events*. That is, each of these branches can only be arrived at by traversing the branches leading to it. Probabilistically, this means that the event associated with a branch is conditional on the the events of the branches leading up to it having taken place. Normally, the event tree is labeled with both the events and their conditional probabilities. In fact, an event tree is of limited use if the probabilities are not labeled on all branches.
In addition, the branches leading from any node must form a partition of the sample space – the sum of probabilities of all branches leaving any node must be 1.0.

Finally, you can only be on one branch at a time (which is really a restatement of the partition requirement). If you have drawn an event tree and realize that one or more of your possible outcomes has you on more than one branch at the same time, then your tree is wrong. In this sense, it sometimes does help to draw a Venn diagram first (corresponding to a particular node) – there should be one branch leaving the node for each separate area on the Venn diagram. For example, a Venn diagram with two overlapping circles corresponding to events A and B would have four branches, one for each of $(A \cap B)$, $(A^c \cap B)$, $(A \cap B^c)$ and $(A^c \cap B^c)$.

Consider the question on the network cards: The network cards must *first be made at a plant*, then *depending on where they were made*, they could be *defective* or not. The event tree for this problem is as follows



Note that there are six 'paths' on the tree. When a network card is selected at random, exactly one of these paths will have been followed. Recall that interest is in finding $\mathrm{P}\big[D\big]$. The event D will have occurred if either the 1st, 3rd, or 5th path was followed. That is, the probability that the 1st, 3rd, or 5th path was followed is sought. If the first path is followed, then the event $A \cap D$ has occurred. The probability that the 1st path was followed is $\mathrm{P}\big[A \cap D\big] = \mathrm{P}\big[D|A\big]\,\mathrm{P}\big[A\big] = 0.01(0.5) = 0.005$. Looking back at the calculation performed in Approach 1, $\mathrm{P}\big[D\big]$ was computed as

$$\begin{aligned}\mathrm{P}\big[D\big] &= \mathrm{P}\big[D|A\big]\,\mathrm{P}\big[A\big] + \mathrm{P}\big[D|B\big]\,\mathrm{P}\big[B\big] + \mathrm{P}\big[D|C\big]\,\mathrm{P}\big[C\big] \\ &= 0.01(0.5) + 0.02(0.3) + 0.03(0.2) \\ &= 0.017 \end{aligned} \tag{19}$$

which, in terms of the event tree, is just the sum of the paths that lead to the outcome that you desire, D. Event trees make 'total probability' problems much simpler. They give a 'picture' of what is going on, and allow the computation of some of the desired probabilities directly.

2.4 Bayes' Theorem

Suppose we want to go the other way, that is, given that event E occurred, what is the probability that event A_i occurred?

We had

$$\mathrm{P}\big[E\big] = \mathrm{P}\big[E \,|\, A_1\big]\,\mathrm{P}\big[A_1\big] + \mathrm{P}\big[E \,|\, A_2\big]\,\mathrm{P}\big[A_2\big] + \cdots \mathrm{P}\big[E \,|\, A_n\big]\,\mathrm{P}\big[A_n\big] \tag{20}$$

and we want $\mathrm{P}\big[A_i \,|\, E\big]$.

Now we know that $\mathrm{P}\big[A_i \cap E\big] = \mathrm{P}\big[E \cap A_i\big]$ so that
$\mathrm{P}\big[A_i \,|\, E\big]\,\mathrm{P}\big[E\big] = \mathrm{P}\big[E \,|\, A_i\big]\,\mathrm{P}\big[A_i\big]$ which we can solve for $\mathrm{P}\big[A_i \,|\, E\big]$,

$$\mathrm{P}\big[A_i \,|\, E\big] = \frac{\mathrm{P}\big[E \,|\, A_i\big]\,\mathrm{P}\big[A_i\big]}{\mathrm{P}\big[E\big]} \tag{21}$$

or, since we know $\mathrm{P}\big[E\big]$,

$$\begin{aligned}\mathrm{P}\big[A_i \,|\, E\big] &= \frac{\mathrm{P}\big[E \,|\, A_i\big]\,\mathrm{P}\big[A_i\big]}{\mathrm{P}\big[E \,|\, A_1\big]\,\mathrm{P}\big[A_1\big] + \cdots + \mathrm{P}\big[E \,|\, A_n\big]\,\mathrm{P}\big[A_n\big]} \\ &= \frac{\mathrm{P}\big[E \,|\, A_i\big]\,\mathrm{P}\big[A_i\big]}{\sum_{j=1}^{n} \mathrm{P}\big[E \,|\, A_j\big]\,\mathrm{P}\big[A_j\big]}\end{aligned} \tag{22}$$

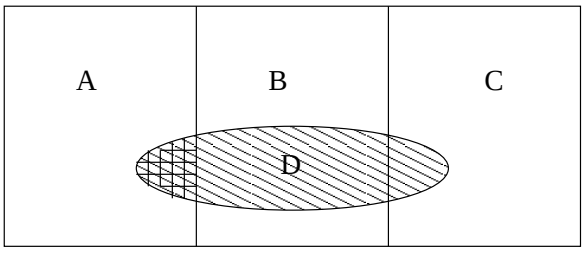
which is Bayes' Theorem.

Return to the manufacturer of network cards from above. If a network card is selected at random and found to be defective, what is the probability that it came from plant A?

Set-up: Same as before, except now the probability of interest is $\mathrm{P}\big[A|D\big]$.

Two possible approaches:

Approach 1
Recall that $\mathrm{P}\big[A|D\big] = \frac{\mathrm{P}[A \cap D]}{\mathrm{P}[D]}$. The relevant quantities are depicted in the following Venn diagram.



where $\mathrm{P}\big[A|D\big]$ can be computed as follows:

$$\begin{aligned}
\mathrm{P}\big[A|D\big] &= \frac{\mathrm{P}\big[A \cap D\big]}{\mathrm{P}\big[D\big]} \\
&= \frac{\mathrm{P}\big[A \cap D\big]}{\mathrm{P}\big[(A \cap D) \cup (B \cap D) \cup (C \cap D)\big]} \\
&= \frac{\mathrm{P}\big[A \cap D\big]}{\mathrm{P}\big[A \cap D\big] + \mathrm{P}\big[B \cap D\big] + \mathrm{P}\big[C \cap D\big]} \\
&\qquad \text{since } A \cap D,\ B \cap D, \text{ and } C \cap D \text{ are mutually exclusive} \\
&= \frac{\mathrm{P}\big[D|A\big]\,\mathrm{P}\big[A\big]}{\mathrm{P}\big[D|A\big]\,\mathrm{P}\big[A\big] + \mathrm{P}\big[D|B\big]\,\mathrm{P}\big[B\big] + \mathrm{P}\big[D|C\big]\,\mathrm{P}\big[C\big]} \\
&= \frac{0.01(0.5)}{0.01(0.5) + 0.02(0.3) + 0.03(0.2)} \\
&= \frac{0.005}{0.017} \\
&= 0.294
\end{aligned} \tag{23}$$

Note that the denominator had already been calculated in the previous question, however the computations have been reproduced here for completeness.

Approach 2
The probability $\mathrm{P}\big[A|D\big]$ can also be easily computed from the event tree. The probability that A has occurred *given that D has occurred* is sought. In terms of the paths on the tree, *we know that one of the 1st, 3rd, or 5th paths has been taken*. Thus, the probability that *the 1st was taken* is just its 'weight' relative to all three possible paths, namely

$$\begin{aligned}
\mathrm{P}\big[A|D\big] &= \frac{\mathrm{P}\big[D|A\big]\,\mathrm{P}\big[A\big]}{\mathrm{P}\big[D|A\big]\,\mathrm{P}\big[A\big] + \mathrm{P}\big[D|B\big]\,\mathrm{P}\big[B\big] + \mathrm{P}\big[D|C\big]\,\mathrm{P}\big[C\big]} \\
&= \frac{0.01(0.5)}{0.01(0.5) + 0.02(0.3) + 0.03(0.2)} \;=\; \frac{0.005}{0.017} \\
&= 0.294
\end{aligned} \tag{24}$$

which is the ratio of the probabilities (weights). This is actually an application of Bayes' Theorem which, of course, agrees with Approach 1.

Bayes' Theorem is useful for revising or updating probabilities as more data and information becomes available. In the previous question on network cards, we had an *initial* probability that a network card was manufactured at plant A: $\mathrm{P}\big[A\big] = 0.5$. This probability is referred to as the prior probability of A. That is, in the absence of any

other information, a network card chosen at random has a probability of having been manufactured at plant A of 0.5. However, if a network card is chosen at random and found to be defective (and thus there is now more information on the network card), then it was computed that the probability that it was manufactured at plant A was reduced to 0.294. This latter probability is referred to as the posterior probability on A. These *Bayesian* quantities have special significance to engineering design and there are many applications.

2.5 Applications to Reliability Theory

The total probability theorem is useful in evaluating the reliability of complex systems because it allows the designer to assign probabilities to individual events before attempting to calculate the system reliability.

For example, the reliability of a system can be found by

1. identifying all possible *causes* of failure: $C_1, C_2, \ldots$
2. estimate the probability of occurrence of each cause: $\mathrm{P}\big[C_1\big], \mathrm{P}\big[C_2\big], \ldots$
3. estimate the probability of failure given that a cause has occurred: $\mathrm{P}\big[F \,|\, C_1\big], \mathrm{P}\big[F \,|\, C_2\big], \ldots$
4. calculate the total probability of failure,
$$\mathrm{P}\big[F\big] = \mathrm{P}\big[F \,|\, C_1\big]\,\mathrm{P}\big[C_1\big] + \mathrm{P}\big[F \,|\, C_2\big]\,\mathrm{P}\big[C_2\big] + \cdots \tag{25}$$
5. calculate the system reliability, F^c: $\mathrm{P}\big[F^c\big] = 1 - \mathrm{P}\big[F\big]$

3 Random Variables

A random variable is a means of identifying events in numerical terms. For example, if the outcome e_1 means that we've selected an apple and e_2 means that we've selected an orange, then we could let $X(e_1) = 1$ and $X(e_2) = 0$. Then $X > 0$ means that we've selected an apple. We can now use mathematics on X, ie. if our fruit picking experiment is repeated n times and $x_1 = X_1(e)$ is the outcome of the first experiment, $x_2 = X_2(e)$ the outcome of the second, etc., then the total number of apples picked is $\sum_{i=1}^{n} x_i$. Note that we could not directly use mathematics on the actual outcomes themselves.

This example illustrates in a rather simple way the primary motivation for the use of random variables – simply so that we can use mathematics. You might notice one other thing in the previous paragraph. After the 'experiment' has taken place and we know what the outcome is, we refer to the lower case, x_i. That is x_i has a known fixed value while X does not. We say that x is a realization of the random variable X.

This is a rather fine distinction, but suffice to say that we can refer to the probability distribution of the random variable X, but not to the probability distribution of x since x is deterministic.

For each outcome e, there is exactly one value of $x = X(e)$, but different values of e may lead to the same x.

In some cases the sample space is already numerical, for example if the quantity of interest is a flood height, then each observation of X is already a number.

3.1 Probability Distributions

Definition: The cumulative distribution function (CDF) of X is defined by

$$F(x) = \mathrm{P}\left[X \leq x\right] \tag{26}$$

Properties of the CDF

1. $0 \leq F(x) \leq 1$
2. $\lim_{x\to\infty} F(x) = 1$
3. $\lim_{x\to-\infty} F(x) = 0$
4. $F(x)$ is non-decreasing
5. $F(x)$ is right continuous (select x; as we approach it from the right, $F(x)$ remains continuous).

Random variables come in two sorts: those which take *discrete* values and those which take *continuous* values. An example of a *discrete random variable* would be the number of students taking this course in any one year. Clearly, this must be a non-negative integer – we cannot have 62.3 students taking the course. The number of students taking the course can be one of the numbers 0, 1, 2, etc.

A *continuous random variable* is one which can take any value on the real line. An example would be a student's height. A randomly selected student can have any height – one might be 1.683674... m in height, another might be 1.683673... m in height. This random variable can take any one of an infinite number of possible values.

Of course, the distinction between continuous and discrete random variables is often blurred. If students heights are only measured to the nearest cm (1.68, 1.69, 1.70, etc.), then the possible set of student heights has been converted to a *discrete* set! In general, any continuous random variable can be converted to a discrete random variable simply by rounding (and ignoring its continuous nature). Conversely, a discrete random variable can be converted (less easily) to a continuous random variable by simply ignoring its discrete nature and assuming that it varies continuously. This is usually only done when the set of possible discrete values is very large.

In the following discussion we will look at both discrete and continuous random variables. In general, all of the common discrete random variables have continuous analogs. Also, probabilities associated with discrete random variables generally involve summations, while probabilities for continuous random variables involve integrations. Recognizing that integrals are simply summations allows the easy translation from continuous to discrete random variables: if you know a formula for a continuous random variable, the corresponding formula for a discrete random variable is obtained simply by replacing integrals with summations.

Discrete Random Variables

Discrete random variables are such that X takes on only discrete values $\{x_1, x_2, \ldots\}$, ie. have a countable number of outcomes (note that countable just means that the outcomes can be numbered $1, 2, \ldots$, however there could still be an infinite number of them).

We saw in an earlier example that in the discrete case,

$$F(x_i) = \mathrm{P}\left[X \leq x_i\right] \;=\; \sum_{x_j \leq x_i} \mathrm{P}\left[X = x_j\right]. \tag{27}$$

We often denote the probability $\mathrm{P}\left[X = x_j\right]$ simply $p(x_j)$ and call this the *probability mass function* (pmf). $p(x_j)$ can be obtained through experimentation in which the frequency of the outcome x_j is measured and normalized by the total number of experiments (frequency interpretation).

The number $p(x_j)$ is a probability, and as such must lie between 0 and 1, inclusive. If we sum up the probabilities associated with each possible x_j, we must get 1. That is, the probability mass function must satisfy

$$\sum_{\text{all } x_j} p(x_j) = 1 \tag{28}$$

Continuous Random Variables

Continuous random variables can take on an infinite number of possible outcomes – generally X takes values from the real line $\mathbb{R}$. Since the probability $\mathrm{P}\left[X = x\right]$ is infinitesimally small we use the *probability density function* to define probabilities;

$$F(x) = \int_{-\infty}^{x} f(\xi)\, d\xi \tag{29}$$

where $f(x)\, dx = \mathrm{P}\left[x < X \leq x + dx\right]$.

NOTE: $f(x)$ is *not* a probability – we call it a *density* because we have to multiply it by a length (area, volume) to get 'mass' or probability.

Conversely, $f(x) = \dfrac{d}{dx}\Big(F(x)\Big)$

Properties: 1) $f(x) \geq 0 \quad \forall x, \qquad 2) \int_{-\infty}^{\infty} f(x)\,dx = 1$

CDF → Probability	PDF → Probability
$\mathrm{P}\big[X \leq x\big] = F(x)$	$\mathrm{P}\big[X \leq x\big] = \int_{-\infty}^{x} f(\xi)\,d\xi$
$\mathrm{P}\big[X > x\big] = 1 - F(x)$	$\mathrm{P}\big[X > x\big] = \int_{x}^{\infty} f(\xi)\,d\xi$
$\mathrm{P}\big[x_1 < X \leq x_2\big] = F(x_2) - F(x_1)$	$\mathrm{P}\big[a < X \leq b\big] = \int_{a}^{b} f(\xi)\,d\xi$

3.2 Main Descriptors of Distributions

Often the exact features of a distribution are unknown. It is convenient to identify key features of a distribution, for example its *mean* and degree of scatter, or *variance*.

Mean

The mean is the most important characteristic of a random value. It tells us the most about a distribution, namely its central tendency. We denote the mean μ and write

$$\begin{aligned} \mu = \mathrm{E}\big[X\big] &= \sum_i x_i \,\mathrm{P}\,\big[X = x_i\big] \;=\; \sum_i x_i\, p(x_i) && \text{for discrete } X \\ &= \int_{-\infty}^{\infty} x\, f(x)\,dx && \text{for continuous } X \end{aligned} \tag{30}$$

where $\mathrm{E}\big[\cdot\big]$ is the *expectation* operator. This is the *first moment* of $f(x)$ about the origin (in analogy to moments of inertia or area). The expectation operator is defined by

$$\mathrm{E}\big[ANYTHING\big] = \int_{-\infty}^{\infty} (anything)\, f(x)\,dx \tag{31}$$

We'll see more of this later.

Note that if X is discrete, μ need not equal any of the possible values of X. For example, in the three coin toss,

$$\mathrm{E}\big[X\big] = \sum_{i=0}^{3} x_i\, p(x_i) \;=\; 0(\tfrac{1}{8}) + 1(\tfrac{3}{8}) + 2(\tfrac{3}{8}) + 3(\tfrac{1}{8}) \;=\; 1.5 \tag{32}$$

that is, on average 1.5 heads will turn up (just half the number of coin tosses).

Variance

The mean or expected value of the r.v. X tells where the probability distribution is "centered". But is the distribution "skinny", "fat", or somewhere in between? This

distribution "dispersion" is measured by a quantity called the variance of X. This is the second most important characteristic of a random value, namely the degree of scatter or variance which we denote as σ^2. This gives us some additional information about a distribution – it is the second most important piece of information about the distribution. We compute the variance as follows;

$$\begin{aligned}\sigma^2 = \mathrm{Var}\big[X\big] = \sum_i (x_i - \mu)^2 \,\mathrm{P}\,\big[X = x_i\big] \;&=\; \sum_i (x_i - \mu)^2\, p(x_i) && \text{for discrete } X \\ &= \int_{-\infty}^{\infty} (x-\mu)^2\, f(x)\, dx && \text{for continuous } X \quad (33)\end{aligned}$$

This is the second moment of $f(x)$ about the mean.

We call $\sigma = \sqrt{\sigma^2}$ the standard deviation (which has the same units as X and μ). The dimensionless coefficient of variation, v, is defined as $\frac{\sigma}{\mu}$ and its size gives us a direct sense of how variable the random variable is. For example, a random variable with mean 1 and standard deviation 3 is highly variable whereas a random variable with mean $1,000,000$ and standard deviation 3 is pretty well a constant.

Making use of the definition for μ we could write

$$\begin{aligned}\mathrm{Var}\big[X\big] &= \sum_i x_i^2\, p(x_i) - \mu^2 && \text{for discrete } X \\ &= \int_{-\infty}^{\infty} x^2\, f(x)\, dx - \mu^2 && \text{for continuous } X \\ &= \mathrm{E}\big[X^2\big] - \mathrm{E}^2\big[X\big] && \text{for both} \quad (34)\end{aligned}$$

Expectations

In general, the expectation of $g(X)$ which can be any function of X is

$$\begin{aligned}\mathrm{E}\big[g(X)\big] &= \sum_i g(x_i)\, p(x_i) && \text{for discrete } X \\ &= \int_{-\infty}^{\infty} g(x)\, f(x)\, dx && \text{for continuous } X \quad (35)\end{aligned}$$

so we see that

$$\begin{aligned}\mu = \mathrm{E}\big[X\big] &= \int_{-\infty}^{\infty} x\, f(x)\, dx \\ \sigma^2 = \mathrm{E}\big[(X-\mu)^2\big] &= \int_{-\infty}^{\infty} (x-\mu)^2\, f(x)\, dx \;=\; \mathrm{E}\big[X^2\big] - \mathrm{E}^2\big[X\big] \qquad (36)\end{aligned}$$

$$\therefore \quad \begin{array}{lll} \mathrm{E}\big[X\big] & \rightarrow & \mu \\ \mathrm{E}\big[X^2\big] & \rightarrow & \sigma^2 \\ \cdot & & \\ \cdot & & \end{array}$$

$\mathrm{E}\big[X^n\big]$ is the n^{th} moment of the pdf about the origin.

Comments:

1. $\mathrm{E}\big[\cdot\big]$ is a linear operator (i.e. integration is a linear operation because it is really just a summation). Thus if $g(X)$ is a linear function, $g(X) = a + bX$, then

$$\mathrm{E}\big[g(X)\big] = \mathrm{E}\big[a + bX\big] \;=\; a + b\mathrm{E}\big[X\big] \tag{37}$$

If $g(X)$ is not linear, then you must use the integral form to find expectations, for example if $g(X) = aX + bX^2 + c\sin(X)$ then

$$\mathrm{E}\big[g(X)\big] = \mathrm{E}\big[aX + bX^2 + c\sin(X)\big] \;=\; a\mathrm{E}\big[X\big] + b\mathrm{E}\big[X^2\big] + c\mathrm{E}\big[\sin(X)\big] \tag{38}$$

where $\mathrm{E}\big[\sin(X)\big] = \displaystyle\int_{-\infty}^{\infty} \sin(x)\, f(x)\, dx$

$$\mathrm{E}\big[X^2\big] = \int_{-\infty}^{\infty} x^2\, f(x)\, dx$$

$$\mathrm{E}\big[X\big] = \int_{-\infty}^{\infty} x\, f(x)\, dx$$

2. The variance of a linear function $g(X) = a + bX$ works out to be reasonably simple

$$\begin{aligned} \mathrm{Var}\big[a + bX\big] &= \mathrm{E}\Big[\big\{(a + bX) - \mathrm{E}\big[a + bX\big]\big\}^2\Big] \\ &= \mathrm{E}\Big[\big\{a + bX - a - b\mathrm{E}\big[X\big]\big\}^2\Big] \\ &= \mathrm{E}\Big[\big\{b(X - \mathrm{E}\big[X\big])\big\}^2\Big] \\ &= b^2\mathrm{E}\big[(X - \mathrm{E}\big[X\big])^2\big] \\ &= b^2\mathrm{Var}\big[X\big] \end{aligned} \tag{39}$$

Higher Order Moments

Skewness: degree of asymmetry

$$\begin{aligned} \theta = \frac{\mathrm{E}\big[(X-\mu)^3\big]}{\sigma^3} &= \tfrac{1}{\sigma^3}\textstyle\sum_i (x_i - \mu)^3\, p(x_i) \\ &= \tfrac{1}{\sigma^3}\textstyle\int_{-\infty}^{\infty} (x-\mu)^3\, f(x)\, dx \end{aligned} \tag{40}$$

If the distribution is symmetric about its mean, then $\theta = 0$ (in fact all odd moments about the mean are zero).

3.3 Covariance and Correlation

Often one must consider more than one random variable at a time. For example, the friction angle and cohesion in the soil at a point are random variables. These two soil properties can be modeled by two random variables, and since they likely influence one another (or they are jointly influenced by some other factor), they must be characterized by a *bivariate distribution*.

Properties of the Bivariate Distribution

Discrete:

a) $f_{XY}(x,y) = \mathrm{P}\big[X = x \cap Y = y\big]$

b) $0 \le f_{XY}(x,y) \le 1$

c) $\sum\limits_{\text{all } x} \sum\limits_{\text{all } y} f_{XY}(x,y) = 1$

Continuous:

a) $f_{XY}(x,y) \ge 0 \quad$ for all $(x,y) \in \Re^2$

b) $\displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{XY}(x,y)\,dx\,dy = 1$

c) $\mathrm{P}\big[x_1 < X \le x_2 \cap y_1 < Y \le y_2\big] = \displaystyle\int_{y_1}^{y_2}\int_{x_1}^{x_2} f_{XY}(x,y)\,dx\,dy$

Consider now two random variables, which we will call X and Y. Recall that the first two primary characteristics of the distributions governing X and Y are

$$\begin{aligned} \mu_X &= \mathrm{E}\big[X\big]\,, & \sigma_X^2 &= \mathrm{E}\big[(X-\mu_X)^2\big] \\ \mu_Y &= \mathrm{E}\big[Y\big]\,, & \sigma_Y^2 &= \mathrm{E}\big[(Y-\mu_Y)^2\big] \end{aligned} \tag{41}$$

which are obtained using the distributions of X and Y respectively. Sometimes the value that X takes on has absolutely no affect on the value that Y takes on. In this case, we say that the random variables, X and Y, are independent. In general, however, X does affect the value that Y takes on (for example, if X is temperature, and Y is ice-cream sales, then an increase in X is expected to cause an increase in Y). The primary characteristic reflecting the degree of linear dependence between X and Y is their *covariance*, which is defined as follows;

$$\begin{aligned} \mathrm{Cov}\big[X,Y\big] &= \mathrm{E}\big[(X-\mu_X)(Y-\mu_Y)\big] \;=\; \mathrm{E}\big[XY - \mu_Y X - \mu_X Y + \mu_X\mu_Y\big] \\ &= \mathrm{E}\big[XY\big] - \mathrm{E}\big[X\big]\,\mathrm{E}\big[Y\big] \end{aligned} \tag{42}$$

Note that $\mathrm{Cov}\big[X,X\big] = \mathrm{Var}\big[X\big]$, so that the covariance is a second moment. In fact $\mathrm{E}\big[XY\big]$ is the joint second moment of X and Y,

$$\mathrm{E}\big[XY\big] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy f_{XY}(x,y)\,dx\,dy \tag{43}$$

If X is independent of Y (which we could write as $X \perp Y$), then $f_{XY}(x,y) = f_X(x)\cdot f_Y(y)$ so that

$$\begin{aligned}\mathrm{E}\big[XY\big] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy f_X(x) f_Y(y)\,dx\,dy = \int_{-\infty}^{\infty} x f_X(x)\,dx \int_{-\infty}^{\infty} y f_Y(y)\,dy \\ &= \mathrm{E}\big[X\big]\,\mathrm{E}\big[Y\big]\end{aligned} \tag{44}$$

in which case $\mathrm{Cov}\big[X,Y\big] = 0$. The converse is not necessarily true, that is if $\mathrm{Cov}\big[X,Y\big] = 0$, X and Y are not necessarily independent – we can only say that they are *uncorrelated*.

The covariance measures the degree of *linear* dependence between X and Y. However, the magnitude of the covariance is not intuitively meaningful since it depends on the variability of the two random variables. A more meaningful measure of the degree of linear dependence between X and Y is the *correlation coefficient* which is a normalized quantity

$$\rho_{XY} = \frac{\mathrm{Cov}\big[X,Y\big]}{\sigma_X \sigma_Y} \tag{45}$$

Again, this will be zero if X and Y are independent or uncorrelated. Note that if $Y = X$, that is X and Y are completely linearly dependent, then $\mathrm{Cov}\big[X,Y\big] = \mathrm{Cov}\big[X,X\big] = \sigma_X^2$ and we get $\rho_{XY} = 1$. In fact we can show that

$$-1 \le \rho_{XY} \le 1 \tag{46}$$

for any X and Y.

3.4 Linear Combinations

If the random variable Y is formed from a linear combination of two other random variables, X_1 and X_2, as in $Y = a_1X_1 + a_2X_2$ then the moments of Y are given by

$$\begin{aligned}\mathrm{E}\big[Y\big] &= a_1\mathrm{E}\big[X_1\big] + a_2\mathrm{E}\big[X_2\big] \\ \mathrm{Var}\big[Y\big] &= a_1^2\mathrm{Var}\big[X_1\big] + a_2^2\mathrm{Var}\big[X_2\big] + 2a_1a_2\mathrm{Cov}\big[X_1,X_2\big]\end{aligned} \tag{47}$$

In the general case if $Y = \sum_{i=1}^{n} a_i X_i$, then

$$\begin{aligned} \mathrm{E}\big[Y\big] &= \sum_{i=1}^{n} a_i \mathrm{E}\big[X_i\big] \\ \mathrm{Var}\big[Y\big] &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \mathrm{Cov}\big[X_i, X_j\big] \end{aligned} \tag{48}$$

Also if we have $Y = \sum_{i=1}^{n} a_i X_i, \quad Z = \sum_{j=1}^{m} b_j X_j$ then

$$\mathrm{Cov}\big[Y, Z\big] = \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j \mathrm{Cov}\big[X_i, X_j\big] \tag{49}$$

3.5 Common Discrete Distributions

3.5.1 Bernoulli Trials

If each of a sequence of trials has two possible outcomes, $S_j = \{S, F\}$ where S_j is the sample space of the j^{th} trial, and the trials are independent with constant probability of success ($p = \mathrm{P}\big[S\big]$), then the sequence of trials is called a *Bernoulli Process*. There are many examples of Bernoulli processes: one might model the failures of individual telescopes in a large array of radio telescopes using a Bernoulli process. Students passing or failing a course might also constitute a Bernoulli process (if they work independently!).

If we let

$$X_j = \begin{cases} 1 & \text{if the } j^{th} \text{ trial results in } \{S\}, \\ 0 & \text{if the } j^{th} \text{ trial results in } \{F\}. \end{cases} \tag{50}$$

then the Bernoulli distribution is given by

$$\begin{aligned} \mathrm{P}\big[X_j = 1\big] &= p \\ \mathrm{P}\big[X_j = 0\big] &= 1 - p = q \end{aligned} \tag{51}$$

for all $j = 1, 2, \ldots$. For the individual trials we have the following results

$$\begin{aligned}
\mathrm{P}\big[X_1 = x_1 \cap X_2 = x_2 \cap \cdots X_n = x_n\big] &= \prod_{i=1}^{n} \mathrm{P}\big[X_i = x_i\big] \\
\mathrm{E}\big[X_j\big] &= \sum_{i=0}^{1} i\mathrm{P}\big[X_j = i\big] = 0(1-p) + 1(p) \;=\; p \\
\mathrm{E}\big[X_j^2\big] &= \sum_{i=0}^{1} i^2\mathrm{P}\big[X_j = i\big] = 0^2(1-p) + 1^2(p) \;=\; p \\
\mathrm{Var}\big[X_j\big] &= \mathrm{E}\big[X_j^2\big] - \mathrm{E}^2\big[X_j\big] = p - p^2 \;=\; pq
\end{aligned} \tag{52}$$

3.5.2 Binomial Distribution

Now let

$$N_n = X_1 + X_2 + \cdots X_n \tag{53}$$

denote the number of successes amongst n Bernoulli trials.

In general N_n follows the binomial distribution with

$$\mathrm{P}\big[N_n = x\big] = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, 1, 2, \ldots, n \tag{54}$$

where $p^x(1-p)^{n-x}$ is the probability of obtaining a particular sequence of x successes in n trials and $\binom{n}{x}$ is the number of possible outcomes having x successes in n trials.

The expected number of successes in n trials can be found directly from the definition of expectation (discrete case),

$$\begin{aligned}
\mathrm{E}\big[N_n\big] &= \sum_{x=0}^{n} x \binom{n}{x} p^x q^{n-x} = \sum_{x=0}^{n} x \left(\frac{n!}{x!(n-x)!}\right) p^x q^{n-x} \\
&= np \sum_{x=1}^{n} \frac{(n-1)!}{(x-1)!(n-x)!} p^{x-1} q^{n-x} \\
&= np \sum_{x=0}^{(n-1)} \frac{(n-1)!}{x!((n-1)-x)!} p^x q^{(n-1)-x} = np(p+q)^{n-1} \\
\therefore\ \mathrm{E}\big[N_n\big] &= np
\end{aligned} \tag{55}$$

Alternatively, we could write

$$\begin{aligned}\mathrm{E}\big[N_n\big] &= \mathrm{E}\big[X_1 + X_2 + \cdots + X_n\big] \\ &= \mathrm{E}\big[X_1\big] + \mathrm{E}\big[X_2\big] + \cdots + \mathrm{E}\big[X_n\big] \\ &= np \end{aligned} \tag{56}$$

To find the variance of N_n, we need first to find

$$\begin{aligned}\mathrm{E}\big[N_n^2\big] &= \sum_{x=0}^{n} x^2 \binom{n}{x} p^x q^{n-x} = \sum_{x=1}^{n} x^2 \left(\frac{n!}{x!(n-x)!}\right) p^x q^{n-x} \\ &= np \sum_{x=1}^{n} x \left(\frac{(n-1)!}{(x-1)!(n-x)!}\right) p^{x-1} q^{n-x} \\ &= np \sum_{x=0}^{n-1} (x+1) \left(\frac{(n-1)!}{x!(n-1-x)!}\right) p^{x} q^{n-1-x} \\ &= np\,\{(n-1)p + 1\} \\ \therefore\ \mathrm{Var}\big[N_n\big] &= \mathrm{E}\big[N_n^2\big] - \mathrm{E}^2\big[N_n\big] = npq \end{aligned} \tag{57}$$

The same result could have been (much more easily) obtained by considering the variance of a sum of independent random variables.

3.5.3 Geometric Distribution

Consider a Bernoulli process in which T_1 is the number of trials ("time") required to achieve the first success. Thus if $T_1 = 3$, then we must have had 2 failures followed by a success (the value of T_1 fully prescribes the sequence of trials). This has probability

$$\mathrm{P}\big[T_1 = 3\big] = \mathrm{P}\big[\{FFS\}\big] = q^2 p \tag{58}$$

In general

$$\mathrm{P}\big[T_1 = k\big] = q^{k-1} p, \qquad k = 1, 2, \ldots \tag{59}$$

Note that this is a valid pmf since

$$\sum_{k=1}^{\infty} q^{k-1} p = p \sum_{k=0}^{\infty} q^k \ = \ \frac{p}{1-q} = 1 \tag{60}$$

Properties:

Mean Recurrence Time (Return Period):

$$\begin{aligned}\mathrm{E}\big[T_1\big] &= \sum_{k=1}^{\infty} kpq^{k-1} = p\sum_{k=1}^{\infty} kq^{k-1} = p\frac{d}{dq}\sum_{k=1}^{\infty} q^k = p\frac{d}{dq}\left(\frac{q}{1-q}\right) \\ &= p\left(\frac{1}{(1-q)^2}\right) \;=\; \frac{1}{p}\end{aligned} \tag{61}$$

Variance:

$$\begin{aligned}\mathrm{E}\big[T_1^2\big] &= \sum_{k=1}^{\infty} k^2 pq^{k-1} = p\sum_{k=1}^{\infty} k^2 q^{k-1} = p\frac{d}{dq}\sum_{k=1}^{\infty} kq^k \\ &= p\frac{d}{dq}\left(\frac{q}{(1-q)^2}\right) = \frac{1}{p} + \frac{2q}{p^2} \\ \therefore\ \mathrm{Var}\big[T_1\big] &= \mathrm{E}\big[T_1^2\big] - \mathrm{E}^2\big[T_1\big] = \frac{q}{p^2}\end{aligned} \tag{62}$$

The geometric distribution is a memoryless process, as is the exponential distribution (which is its continuous counterpart), as we will see. That is for non-negative and integer values of t and k,

$$\begin{aligned}\mathrm{P}\big[T_1 > t+k \,|\, T_1 > t\big] &= \frac{\mathrm{P}\big[T_1 > t+k \cap T_1 > k\big]}{\mathrm{P}\big[T_1 > t\big]} = \frac{\mathrm{P}\big[T_1 > t+k\big]}{\mathrm{P}\big[T_1 > t\big]} \\ &= \frac{\sum_{m=t+k+1}^{\infty} q^{m-1}p}{\sum_{n=t+1}^{\infty} q^{n-1}p} = \frac{\sum_{m=t+k+1}^{\infty} q^{m-1}}{\sum_{n=t+1}^{\infty} q^{n-1}} = \frac{\sum_{m=t+k}^{\infty} q^{m}}{\sum_{n=t}^{\infty} q^{n}} \\ &= \frac{q^k \sum_{m=t+k}^{\infty} q^{m-k}}{\sum_{n=t}^{\infty} q^n} \;=\; q^k\end{aligned} \tag{63}$$

$$\text{but } \mathrm{P}\big[T_1 > k\big] = \sum_{k+1}^{\infty} q^{m-1}p = p\sum_{m=k}^{\infty} q^m = pq^k \sum_{m=0}^{\infty} q^m = pq^k\left(\frac{1}{1-q}\right) = q^k$$

$$\therefore \quad \mathrm{P}\big[T_1 > t+k | T_1 > t\big] = \mathrm{P}\big[T_1 > k\big] \tag{64}$$

3.5.4 Negative Binomial Distribution

Suppose we wish to know the number of trials ("time") of a Bernoulli process until the k'th success. Letting T_k be the number of trials until the k'th success, then

$$\mathrm{P}\big[T_k = m\big] = \binom{m-1}{k-1} p^k q^{m-k} \quad \text{for } m = k, k+1, \ldots \tag{65}$$

Properties

Mean:

$$\begin{aligned}
\mathrm{E}\big[T_k\big] &= \sum_{j=k}^{\infty} j\mathrm{P}\big[T_k = j\big] = \sum_{j=k}^{\infty} j\binom{j-1}{k-1}p^k q^{j-k} \\
&= \sum_{j=k}^{\infty} j\left(\frac{(j-1)!}{(k-1)!(j-k)!}\right)p^k q^{j-k} = kp^k \sum_{j=k}^{\infty}\left(\frac{j!}{k!(j-k)!}\right)q^{j-k} \\
&= kp^k\left[1 + (k+1)q + \frac{(k+2)(k+1)}{2!}q^2 + \frac{(k+3)(k+2)(k+1)}{3!}q^3 + \cdots\right] \\
&= \frac{kp^k}{(1-q)^{k+1}} \;=\; \frac{k}{p}
\end{aligned} \tag{66}$$

Variance: To get the variance, $\mathrm{Var}\big[T_k\big]$, we'll write

$$T_k = T_1 + (T_2 - T_1) + \cdots + (T_k - T_{k-1}) \tag{67}$$

which are independent due to the properties of a Bernoulli sequence. Now $\mathrm{E}\big[T_1\big] = 1/p$ as we saw earlier. Similarly $\mathrm{E}\big[T_j - T_{j-1}\big] = 1/p$ due to memorylessness.

$$\therefore \quad \mathrm{E}\big[T_k\big] = \mathrm{E}\big[T_1\big] + \mathrm{E}\big[T_2 - T_1\big] + \cdots + \mathrm{E}\big[T_k - T_{k-1}\big] \;=\; \frac{k}{p} \tag{68}$$

as found above. Now due to independence of the terms,

$$\begin{aligned}
\mathrm{Var}\big[T_k\big] &= \mathrm{Var}\big[T_1\big] + \mathrm{Var}\big[T_2 - T_1\big] + \cdots + \mathrm{Var}\big[T_k - T_{k-1}\big] \\
&= k\mathrm{Var}\big[T_1\big] \\
&= \frac{kq}{p^2}
\end{aligned} \tag{69}$$

3.5.5 Poisson Distribution

The Poisson distribution governs many 'rate' dependent processes (for example, arrivals of vehicles at an intersection or 'packets' through a computer net). The assumptions on which a Poisson process is based are;

1. events occur at random and at any point in time (or space),
2. the occurrence of an event in a given time (or space) interval is independent of events occurring in disjoint intervals,
3. the probability of an event occurring in a small interval, Δt, is proportional to the size of Δt, ie. is $\lambda\Delta t$, where λ is the mean *rate* of occurrence.

4. for $\Delta t \to 0$, the probability of two or more events in Δt is negligible.

Now let N_t be the number of events occurring in the interval $[0, t]$. We can show that N_t follows the distribution

$$\mathrm{P}\big[N_t = k\big] = \frac{(\lambda t)^k}{k!}\, e^{-\lambda t}, \qquad k = 0, 1, 2, \ldots \tag{70}$$

Properties

Mean:

$$\begin{aligned}\mathrm{E}\big[N_t\big] &= \sum_{j=0}^{\infty} j\, \frac{(\lambda t)^j}{j!}\, \frac{}{e}^{-\lambda t} = \lambda t e^{-\lambda t} \sum_{j=1}^{\infty} \frac{(\lambda t)^{j-1}}{(j-1)!} = \lambda t e^{-\lambda t} \sum_{j=0}^{\infty} \frac{(\lambda t)^j}{j!} \\ &= \lambda t \end{aligned} \tag{71}$$

Variance:

$$\begin{aligned}\mathrm{E}\big[N_t^2\big] &= \sum_{j=0}^{\infty} j^2\, \frac{(\lambda t)^j}{j!}\, e^{-\lambda t} = \lambda t e^{-\lambda t} \sum_{j=0}^{\infty} (j+1) \frac{(\lambda t)^j}{j!} \\ &= \lambda t e^{-\lambda t} \left[\sum_{j=0}^{\infty} j \frac{(\lambda t)^j}{j!} + \sum_{j=0}^{\infty} \frac{(\lambda t)^j}{j!} \right] \\ &= (\lambda t)^2 + (\lambda t) \\ \therefore \quad \mathrm{Var}\big[N_t\big] &= \mathrm{E}\big[N_t^2\big] - \mathrm{E}^2\big[N_t\big] = \lambda t \end{aligned} \tag{72}$$

Comparison to Binomial Distribution

Suppose that a systems engineer analyses a network and verifies that an average of 60 packets of information pass through a gateway per hour. What is the probability that 10 packets will pass through in a 10-minute interval?

Solution:

Divide the hour into 120 30-second intervals and assume that no more than 1 packet can be transmitted in a 30-second time interval. Thus $p = 60/120 = 0.5$ is the probability of 1 packet arriving in a 'trial' interval, and

$$\mathrm{P}\big[\text{10 packets in 10 minutes}\big] \simeq \binom{20}{10} (0.5)^{10} (0.5)^{20-10} = 0.176 \tag{73}$$

Of course, two or more packets could be transmitted in 30 seconds (hopefully!). An improved solution is obtained using a shorter 'trial' interval, ie. if 10-second intervals

were used then $p = 60/360 = \frac{1}{6}$ and

$$\mathrm{P}\big[\text{10 packets in 10 minutes}\big] \simeq \binom{60}{10}(\tfrac{1}{6})^{10}(\tfrac{5}{6})^{50} = 0.137 \tag{74}$$

In general, if time t is divided into n intervals then $p = \frac{\lambda t}{n}$ and

$$\mathrm{P}\big[N_t = k\big] = \binom{n}{k}\left(\frac{\lambda t}{n}\right)^k\left(1-\frac{\lambda t}{n}\right)^{n-k} \tag{75}$$

where λt is the mean number of events occurring in the time t. Now if packets pass the gate 'instantaneously' and can arrive at any time then

$$\begin{aligned}\mathrm{P}\big[N_t = k\big] &= \lim_{n\to\infty}\binom{n}{k}\left(\frac{\lambda t}{n}\right)^k\left(1-\frac{\lambda t}{n}\right)^{n-k} \\ &= \lim_{n\to\infty}\left[\left\{\frac{n}{n}\cdot\frac{n-1}{n}\cdots\frac{n-k+1}{n}\right\}\frac{(\lambda t)^k}{k!}\left(1-\frac{\lambda t}{n}\right)^n\left(1-\frac{\lambda t}{n}\right)^{-k}\right]\end{aligned} \tag{76}$$

but since $\lim_{n\to\infty}\left(1-\frac{\lambda t}{n}\right)^n = e^{-\lambda t}$ this reduces to

$$\mathrm{P}\big[N_t = k\big] = \frac{(\lambda t)^k}{k!}\,e^{-\lambda t} \tag{77}$$

which is the Poisson distribution. Thus we see that the Poisson distribution is a limiting case of the Binomial distribution.

For our problem $\lambda = 1$ packet/minute and $t = 10$ minutes so that, using the Poisson distribution

$$\mathrm{P}\big[X_{10} = 10\big] = \frac{(10)^{10}}{10!}\,e^{-10} \;=\; 0.125.$$

3.6 Common Continuous Distributions

3.6.1 Exponential Distribution

The exponential distribution is useful to describe 'time-to-failure' type problems. It also governs the time between occurrences of a Poisson process. If T is the time to the occurrence (or failure) in question, then

$$f(t) = \begin{cases}\lambda e^{-\lambda t} & \text{if } t \geq 0 \\ 0 & \text{otherwise}\end{cases} \tag{78}$$

where λ is the *mean rate* of occurrence (or failure).

Characteristics:

$$\begin{aligned} \mathrm{E}\big[T\big] &= \frac{1}{\lambda} \\ \mathrm{Var}\big[T\big] &= \frac{1}{\lambda^2} \end{aligned} \tag{79}$$

Memoryless Property:

Let T denote the time between detections of a rare particle at a geiger counter. Assume that T has an exponential distribution with a mean of 4 minutes. Thus, $T \sim \exp\left(\frac{1}{\lambda} = 4 \rightarrow \lambda = 0.25\right)$. The probability that a particle is detected within 30 seconds of starting the counter is

$$\mathrm{P}\big[T < 30 \text{ s}\big] = \mathrm{P}\big[T < 0.5 \text{ min}\big] = 1 - e^{-0.5 \times 0.25} = 0.1175 \tag{80}$$

Now, suppose that the geiger counter is turned on and 3 minutes pass without detection of a particle. What is the probability that a particle will be detected in the next 30 seconds? Because 3 minutes have gone by without detection, you might feel that a detection is "due". That is, that the probability of detection in the next 30 seconds should be greater than 0.1175. However, for the exponential distribution, this is not true. In fact,

$$\begin{aligned} \mathrm{P}\big[T < 3.5 | T > 3\big] &= \frac{\mathrm{P}\big[3 < T < 3.5\big]}{\mathrm{P}\big[T > 3\big]} = \frac{(1 - e^{-3.5 \times 0.25}) - (1 - e^{-3 \times 0.25})}{e^{-3 \times 0.25}} \\ &= 0.1175 \end{aligned} \tag{81}$$

Thus, after waiting for 3 minutes without detection, the probability of detection in the next 30 seconds is the same as the probability of detection in the 30 seconds immediately after starting the counter.

More generally, if $T \sim \exp(\lambda)$ then

$$\begin{aligned} \mathrm{P}\big[T > t + s | T > t\big] &= \frac{\mathrm{P}\big[T > t + s \cap T > t\big]}{\mathrm{P}\big[T > t\big]} = \frac{\mathrm{P}\big[T > t + s\big]}{\mathrm{P}\big[T > t\big]} = \frac{e^{-\lambda(t+s)}}{e^{-\lambda t}} \\ &= e^{-\lambda s} = \mathrm{P}\big[T > s\big] \end{aligned} \tag{82}$$

The exponential distribution is the only continuous probability distribution function to have the memoryless property (note that the discrete geometric distribution, which is the analog of the exponential, also has this property). This is due to the fact that the exponential distribution arises from an infinite number of Bernoulli trials, one for each instant in time (or space) and all of which are independent with constant probability (rate) of "success".

Link to Poisson: It was mentioned earlier that the exponential distribution governs the time between the occurrences of a Poisson process. We can see this as follows;

Let N_t be a Poisson process with mean rate λ. If we start watching this process at time $t = 0$, and we let T be the time until the first "arrival" in this Poisson process, then the event that $T > t$ is the same as the event that no arrivals occur in the time interval $[0, 1]$. Thus,

$$\mathrm{P}\big[T > t\big] = \mathrm{P}\big[N_t = 0\big] = \frac{(\lambda t)^0}{0!}\, e^{\lambda t} = e^{-\lambda t} \tag{83}$$

Thus,

$$F(t) = \mathrm{P}\big[T \leq t\big] = 1 - e^{-\lambda t} \tag{84}$$

But $F(t) = 1 - e^{-\lambda t}$ is the cumulative distribution for the exponential probability density function $f(t) = \lambda e^{-\lambda t}$. This means that T must follow an exponential distribution with parameter λ equal to the Poisson rate. Due to independence between all trials (one at each instant in time) in a Poisson process, the time between any arrivals follows the same exponential distribution.

3.6.2 Gamma Distribution

If T_k is the sum of k independent exponentially distributed random variables E_i, each with parameter λ, that is $T_k = E_1 + E_2 + \cdots E_k$, then

$$f_{T_k}(t) = \begin{cases} \frac{\lambda}{\Gamma(k)} (\lambda t)^{k-1}\, e^{-\lambda t} & \text{if } t \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{85}$$

where $\Gamma(\cdot)$ is the Gamma function. For integers, $\Gamma(k) = (k-1)!$. Note that $k = 1$ gives the exponential distribution, as expected.

Link to Poisson: The Gamma distribution governs the time between every k^{th} "arrival" in a Poisson process. That is, if T_k is the time to the k^{rmth} arrival in a Poission process, then T_k is the sum of k independent exponentially distributed "inter-arrival" times and T_k follows a Gamma distribution.

Characteristics:

$$\begin{aligned} \mathrm{E}\big[T_k\big] &= \frac{k}{\lambda} \qquad \Big(= k\mathrm{E}\big[E_i\big]\Big) \\ \mathrm{Var}\big[T_k\big] &= \frac{k}{\lambda^2} \qquad \Big(= k\mathrm{Var}\big[E_i\big]\Big) \end{aligned} \tag{86}$$

3.6.3 Weibull Distribution

Often, engineers are concerned with the strength properties of materials and the lifetime of manufactured devices. The Weibull distribution has become extremely popular in describing such behavior. Its probability density function, mean, and variance are a

bit clumsy to describe so the cumulative distribution (since this is really all one needs to compute probabilities anyways) will be used.

If a continuous random variable X has a Weibull distribution with parameters $\lambda > 0$ and $\beta > 0$, then it has probability density function

$$f(x) = \frac{\beta}{x}(\lambda x)^{\beta}\, e^{-(\lambda x)^{\beta}}, \qquad \text{for } x > 0 \tag{87}$$

and zero otherwise. The Weibull has a particularly simple cumulative distribution function

$$F(x) = 1 - e^{-(\lambda x)^{\beta}} \qquad \text{for } x \geq 0 \tag{88}$$

Note that the exponential distribution is a special case of the Weibull distribution (set $\beta = 1$). The exponential distribution has constant failure rate whereas the Weibull allows a failure rate that decreases with time (i.e., a system which improves with time; $\beta < 1$) or a failure rate that increases with time (i.e., a system which degrades with time; $\beta > 1$).

The mean and variance of a Weibull distributed random variable are

$$\begin{aligned} \mu &= \frac{1}{\lambda}\,\Gamma\left(1 + \frac{1}{\beta}\right) \\ \sigma^2 &= \frac{1}{\lambda^2}\left[\Gamma\left(1 + \frac{2}{\beta}\right) - \Gamma^2\left(1 + \frac{1}{\beta}\right)\right] \end{aligned} \tag{89}$$

where $\Gamma(x)$ is the Gamma function defined as

$$\Gamma(x) = \int_0^{\infty} e^{-t}\, t^{x-1}\, dt \tag{90}$$

If x is a positive integer, $\Gamma(x) = (x-1)!$.

3.6.4 Uniform Distribution

The uniform distribution is the simplest of distributions. Its general definition is

$$f(x) = \begin{cases} \frac{1}{\beta - \alpha} & \text{if } \alpha \leq x \leq \beta \\ 0 & \text{otherwise} \end{cases} \tag{91}$$

Characteristics:

$$\begin{aligned} \mathrm{E}\big[X\big] &= \int_{\alpha}^{\beta} \frac{x\, dx}{\beta - \alpha} \;=\; \tfrac{1}{2}(\alpha + \beta) \qquad \text{(this is the midpoint)} \\ \mathrm{Var}\big[X\big] &= \int_{\alpha}^{\beta} \frac{x^2\, dx}{\beta - \alpha} - \mathrm{E}^2\big[X\big] \;=\; \tfrac{1}{12}(\beta - \alpha)^2 \end{aligned} \tag{92}$$

3.6.5 Normal Distribution

The normal distribution is perhaps the single most important distribution. This is perhaps because the central limit theorem predicts that many natural 'additive' type phenomena (or phenomena involving many factors) tend towards a normal distribution. We'll look at the central limit theorem shortly.

We say that a random variable X follows a normal (or *Gaussian*) distribution if

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \qquad \text{on } -\infty < x < \infty \tag{93}$$

We also write this as $X \sim N(\mu, \sigma^2)$, where $N(\cdot)$ stands for normal distribution and (μ, σ^2) are the parameters of the distribution.

Properties:

1. symmetric about the mean μ,
2. the maximum point of the distribution occurs at μ (we call this the *mode*)
3. inflection points of $f(x)$ are at $x = \mu \pm \sigma$

Characteristics:

$$\mathrm{E}\big[X\big] = \mu$$
$$\mathrm{Var}\big[X\big] = \sigma^2$$
$$F(x) = \mathrm{P}\big[X \le x\big] = \mathrm{P}\left[Z \le \frac{x-\mu}{\sigma}\right] \;=\; \int_{-\infty}^{\frac{x-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}\, dz \;=\; \Phi\left(\frac{x-\mu}{\sigma}\right) \tag{94}$$

We call $Z = \frac{x-\mu}{\sigma}$ the *standardized normal* variate.

Standard Normal Distribution

If we can transform $X \sim N(\mu, \sigma^2)$ into $Z \sim N(0, 1)$, then we only need one set of probability tables, where

$$f_z(z) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}z^2}$$
$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}\xi^2}\, d\xi \tag{95}$$

Most probability and statistics books have tables of $\Phi(z)$. For any normally distributed X, set

$$Z = \frac{X-\mu}{\sigma} \qquad \Longrightarrow \qquad z = \frac{x-\mu}{\sigma} \tag{96}$$

and use the standard tables available in most probability text books.

One of the nice features about the normal distribution is that the sum of normally distributed random variables remains normally distributed. This is one of the results of the Central Limit Theorem to be considered next. That is, if X and Y are normally distributed, then

$$Z = X - Y \tag{97}$$

is also normally distributed with mean $\mu_Z = \mu_X - \mu_Y$ and variance $\sigma_Z^2 = \sigma_X^2 + \sigma_Y^2$, in the event that X and Y are uncorrelated. If X and Y are not uncorrelated, then see the results presented below for more general results.

The Central Limit Theorem

If $X_1, X_2, \ldots, X_n$ are independent random variables having arbitrary distributions, then the random variable

$$Y = X_1 + X_2 + \cdots + X_n \tag{98}$$

has a normal distribution as $n \to \infty$ if all the X's have about the same 'weight'.

The central limit theorem can be illustrated by considering the binomial distributed random variable

$$N_n = \sum_{i=1}^{n} X_i \tag{99}$$

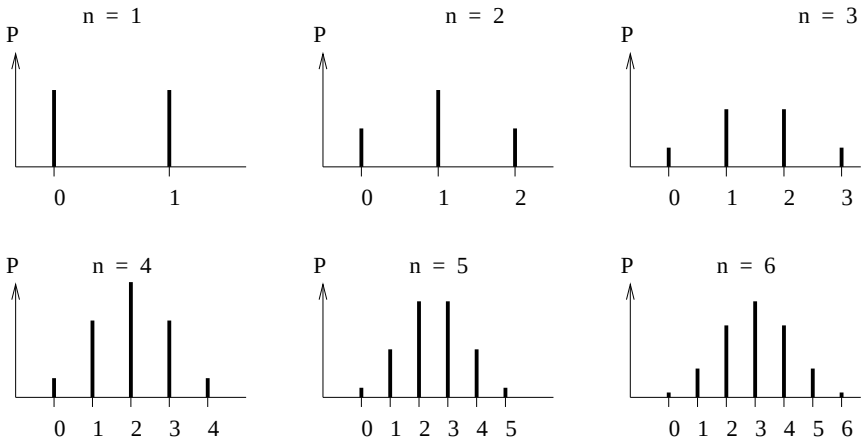
where X_i is a Bernoulli random variable, having value $X_i = 1$ with probability p, and value $X_i = 0$ with probability $q = 1 - p$. Clearly, N_n is a sum of n random variables each of which has a distribution which is distinctly non-normal (as can be seen by the first plot in the following figure). We saw in Chapter 1 that N_n follows a binomial distribution with

$$\mathrm{P}\big[N_n = k\big] = \binom{n}{k} p^k q^{n-k} \tag{100}$$

For $n = 1, 2, \cdots 6$ and $p = 0.5$ we get the following probabilities;

n	k						
	0	1	2	3	4	5	6
1	1/2	1/2					
2	1/4	1/2	1/4				
3	1/8	3/8	3/8	1/8			
4	1/16	1/4	3/8	1/4	1/16		
5	1/32	5/32	5/16	5/16	5/32	1/32	
6	1/64	3/32	15/64	5/16	15/64	3/32	1/64

which have the following bar plots;



Clearly, the distribution for $n = 1$ is about as non-normal as possible (this is the Bernoulli distribution). However, it is clear that as more and more random variables are added together to give N_n, the distribution becomes increasingly normal in shape. The central limit theorem starts to become evident even for $n = 2$.

Specifically, the central limit theorem tells us that if

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{101}$$

where $X_1, X_2, \ldots, X_n$ form a *random sample* from X having mean μ and variance σ^2 (any distribution), then

$$\lim_{n\to\infty} \mathrm{P}\left[\frac{(\bar{X}_n - \mu)}{\sigma/\sqrt{n}} \leq x\right] = \Phi(x) \tag{102}$$

In other words, $\bar{X}$ will tend to be normally distributed for large n.

Implications:

1. the sum of normal variates is normal (for any n) as discussed above,
2. if the distributions of the X's are well-behaved (almost normal), then $n \geq 4$ gives a good approximation to the normal distribution,
3. if the distributions of the X's are uniform (or almost so), then $n \geq 6$ yields a reasonably good approximation to the normal distribution,
4. for poorly-behaved distributions, you may need $n > 100$ before the distribution begins to look normal. This happens, for example, with distributions whose tails do not fall off rapidly.
5. in general, if $n \geq 30$ the sum is deemed to be very closely normally distributed.

Thus for n sufficiently large and $X_1, X_2, \ldots, X_n$ independent and identically distributed (iid), then

$$Y = X_1 + X_2 + \cdots + X_n \tag{103}$$

is approximately normally distributed with

$$\begin{aligned} \mu_Y &= \mathrm{E}\big[Y\big] = n\mathrm{E}\big[X_i\big] \\ \sigma_Y^2 &= \mathrm{Var}\big[Y\big] = n\mathrm{Var}\big[X_i\big] \end{aligned} \tag{104}$$

If the X's are *not* identically distributed, then

$$\begin{aligned} \mu_Y &= \sum_{i=1}^{n} \mathrm{E}\big[X_i\big] \\ \sigma_Y^2 &= \sum_{i=1}^{n} \mathrm{Var}\big[X_i\big] \end{aligned} \tag{105}$$

Normal Approximation to the Binomial

By virtue of the Central Limit Theorem, the binomial distribution, which as you will recall arises from the sum of a sequence of Bernoulli random variables, can be approximated by the normal distribution. Specifically, if N_n is the number of successes in n trials, then

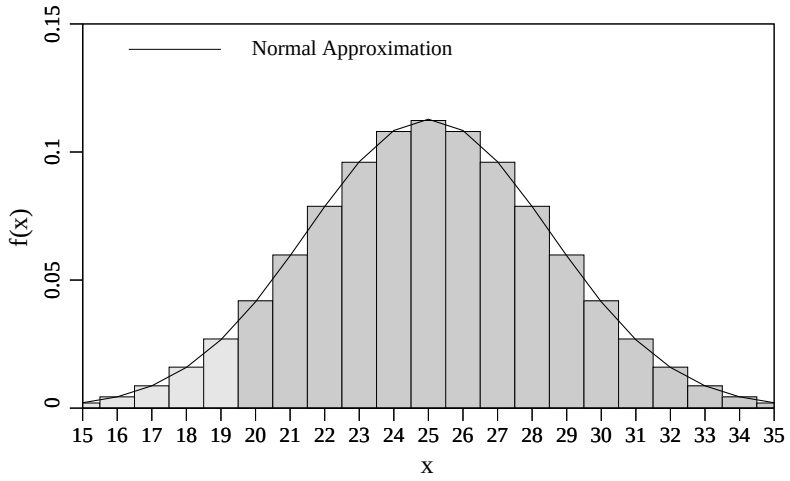
$$N_n = \sum_{i=1}^{n} X_i \tag{106}$$

where X_i is the outcome of a Bernoulli trial ($X_i = 1$ with probability p, $X_i = 0$ with probability $q = 1 - p$). Since N_n is the sum of identically distributed random variables, then if n is large enough, the Central Limit Theorem says that N_n can be approximated by a normal distribution. We generally consider this approximation to be valid when both $np \geq 5$ and $nq \geq 5$. In this case, the normal distribution approximation is specified by its mean and standard deviation;

$$\begin{aligned} \mu &= np \\ \sigma &= \sqrt{npq} \end{aligned} \tag{107}$$

Of course, we know that N_n is discrete while the normal distribution governs a continuous random variable. When we want to find the approximate probability that N_n is greater than or equal to, say, k, using the normal distribution, we should include all

of the binomial *mass* at k. This means that we should look at the normal probability that $(N_n > k - \frac{1}{2})$. For example, in the following plot, the probability that $N_n \geq 20$ is better captured by the area under the normal distribution above 19.5.



In general, the following corrections apply. Similar corrections apply for two-sided probability calculations.

$$\begin{aligned}
\mathrm{P}\big[N_n \geq k\big] &\simeq 1 - \Phi\left(\frac{k - 0.5 - \mu}{\sigma}\right) \\
\mathrm{P}\big[N_n > k\big] &\simeq 1 - \Phi\left(\frac{k + 0.5 - \mu}{\sigma}\right) \\
\mathrm{P}\big[N_n \leq k\big] &\simeq \Phi\left(\frac{k + 0.5 - \mu}{\sigma}\right) \\
\mathrm{P}\big[N_n < k\big] &\simeq \Phi\left(\frac{k - 0.5 - \mu}{\sigma}\right)
\end{aligned} \tag{108}$$

3.6.6 Lognormal Distribution

The random variable X is lognormally distributed if $\ln(X)$ is normally distributed. X has pdf

$$f(x) = \begin{cases} \frac{1}{x\sigma_{\ln x}\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{\ln x - \mu_{\ln x}}{\sigma_{\ln x}}\right)^2} & \text{if } 0 < x < \infty \\ 0 & \text{otherwise} \end{cases} \tag{109}$$

Note that the two parameters

$$\begin{aligned} \mu_{\ln x} &= \mathrm{E}\big[\ln X\big] \\ \sigma^2_{\ln x} &= \mathrm{Var}\big[\ln X\big] \end{aligned} \tag{110}$$

are the mean and variance of the normally distributed variable $\ln X$.

Probabilities

$$\mathrm{P}\big[a < X \le b\big] = \int_a^b f(x)\,dx \;=\; \int_a^b \frac{1}{x\sigma_{\ln x}\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{\ln x - \mu_{\ln x}}{\sigma_{\ln x}}\right)^2} dx \tag{111}$$

now let $z = \dfrac{\ln x - \mu_{\ln x}}{\sigma_{\ln x}} \qquad \Longrightarrow \qquad dz = \dfrac{dx}{x\sigma_{\ln x}}$ so that

$$\mathrm{P}\big[a < X \le b\big] = \int_{\frac{\ln a - \mu_{\ln x}}{\sigma_{\ln x}}}^{\frac{\ln b - \mu_{\ln x}}{\sigma_{\ln x}}} \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}z^2}\, dz \;=\; \Phi\left(\frac{\ln b - \mu_{\ln x}}{\sigma_{\ln x}}\right) - \Phi\left(\frac{\ln a - \mu_{\ln x}}{\sigma_{\ln x}}\right) \tag{112}$$

Mean and Variance

$$\begin{aligned} \mu_X = \mathrm{E}\big[X\big] &= e^{\mu_{\ln x} + \frac{1}{2}\sigma^2_{\ln x}} \\ \sigma^2_X = \mathrm{Var}\big[X\big] &= \mu_X^2\left(e^{\sigma^2_{\ln x}} - 1\right) \end{aligned} \tag{113}$$

alternatively, if you are given μ_X and σ^2_X, you can obtain the parameters $\mu_{\ln x}$ and $\sigma^2_{\ln x}$ as follows;

$$\begin{aligned} \sigma^2_{\ln x} &= \ln\left(1 + \frac{\sigma^2_X}{\mu^2_X}\right) \\ \mu_{\ln x} &= \ln(\mu_X) - \frac{1}{2}\sigma^2_{\ln x} \end{aligned} \tag{114}$$

Other Characteristics

$$\text{mode} = e^{\mu_{\ln x} - \sigma^2_{\ln x}}$$

$$\text{median} = e^{\mu_{\ln x}} \qquad \Big(\mathrm{P}\big[X \le \text{median}\big] = 0.5\Big)$$

$$\text{mean} = e^{\mu_{\ln x} + \frac{1}{2}\sigma^2_{\ln x}}$$

$$\mathrm{E}\big[X^k\big] = e^{k\mu_{\ln x} + \frac{1}{2}k^2\sigma^2_{\ln x}}$$

Multiplicative Property

If $X = Y_1 Y_2 \cdots Y_n$ and each Y_i are (positive) independent lognormally distributed random variables of about the same 'weight', then

$$\ln X = \ln Y_1 + \ln Y_2 + \cdots + \ln Y_n \tag{115}$$

and since a sum of normals remains normal, $\ln X$ tends to a normal distribution with

$$\begin{aligned} \mu_{\ln x} &= \mu_{\ln y_1} + \mu_{\ln y_2} + \cdots + \mu_{\ln y_n} \\ \sigma^2_{\ln x} &= \sigma^2_{\ln y_1} + \sigma^2_{\ln y_2} + \cdots + \sigma^2_{\ln y_n} \end{aligned} \tag{116}$$

(Note that the last line holds only if the random variables are independent). Thus X tends to a lognormal distribution with parameters $\mu_{\ln x}$ and $\sigma^2_{\ln x}$. In particular, if X is any multiplicative function, ie. say

$$X = \frac{AB}{C} \tag{117}$$

and A, B, and C are independent lognormally distributed random variables, then X is also lognormally distributed with

$$\begin{aligned} \mu_{\ln x} &= \mu_{\ln A} + \mu_{\ln B} - \mu_{\ln C} \\ \sigma^2_{\ln x} &= \sigma^2_{\ln A} + \sigma^2_{\ln B} + \sigma^2_{\ln C} \end{aligned} \tag{118}$$

This is a useful property since it can be used to approximate the distribution of many multiplicative functions.

3.7 Extreme Value Distributions

Most engineering systems fail only when extreme loads occur and failure tends to initiate at the weakest point. Thus, it is of considerable interest to investigate the distribution of extreme values. Consider a sequence of n random variables, $X_1, X_2, \ldots, X_n$. This could, for example, be the sequence of tensile strengths of individual links in a chain, or the sequence of daily average windspeeds, or earthquake intensities, etc. Now define the extremes of this set of random variables as

$$\begin{aligned} Y_n &= \max(X_1, X_2, \ldots, X_n) \\ Y_1 &= \min(X_1, X_2, \ldots, X_n) \end{aligned} \tag{119}$$

so that if X_i is the daily average windspeed, then Y_n is the maximum daily average windspeed over n days. Similarly, if X_i is the tensile strength of the i^{th} link in a chain, then Y_1 is the tensile strength of a chain composed of n links.

3.7.1 Exact Extreme Value Distributions

Now let us examine the behaviour of the maximum, Y_n. We know that if the maximum is less than some number y, then each X_i must also be less than y. That is the event $(Y_n \leq y)$ must be equivalent to the event $(X_1 \leq y \cap X_2 \leq y \cap \cdots \cap X_n \leq y)$. In other words the *exact* distribution of Y_n is,

$$\mathrm{P}\big[Y_n \leq y\big] = \mathrm{P}\big[X_1 \leq y \cap X_2 \leq y \cap \cdots \cap X_n \leq y\big] \tag{120}$$

If it can be further assumed that the X's are independent and identically distributed (IID) (if this is not the case the problem becomes very complex and usually only solved via simulation), then

$$\begin{aligned} F_{Y_n}(y) = \mathrm{P}\big[Y_n \leq y\big] &= \mathrm{P}\big[X_1 \leq y\big]\,\mathrm{P}\big[X_2 \leq y\big]\cdots\mathrm{P}\big[X_n \leq y\big] \\ &= \big[F_X(y)\big]^n \end{aligned} \tag{121}$$

and, taking the derivative,

$$f_{Y_n}(y) = \frac{dF_{Y_n}(y)}{dy} = n\left[F_X(y)\right]^{n-1}\frac{dF_X(y)}{dy} = n\left[F_X(y)\right]^{n-1} f_X(y) \tag{122}$$

Now consider the distribution of Y_1. If we proceed as we did for Y_n, then we would look at the event $Y_1 \leq y$. This event just means that $X_1 \leq y$ or $X_2 \leq y$ or . . ., that is

$$\mathrm{P}\big[Y_1 \leq y\big] = \mathrm{P}\big[X_1 \leq y \cup X_2 \leq y \cup \cdot \cup X_n \leq y\big] \tag{123}$$

which expands into $\binom{n}{1} + \binom{n}{2} + \binom{n}{3} + \ldots + \binom{n}{n}$ terms... in other words a *lot* of terms. A better way to work out this distribution is to look at the complement;

$$\begin{aligned} \mathrm{P}\big[Y_1 > y\big] &= \mathrm{P}\big[X_1 > y \cap X_2 > y \cap \cdots \cap X_n > y\big] \\ &= \mathrm{P}\big[X_1 > y\big]\,\mathrm{P}\big[X_2 > y\big]\cdots\mathrm{P}\big[X_n > y\big] \\ &= \big[1 - F_X(y)\big]^n \end{aligned} \tag{124}$$

and since $\mathrm{P}\big[Y_1 > y\big] = 1 - F_{Y_1}(y)$ we get

$$F_{Y_1}(y) = 1 - \big[1 - F_X(y)\big]^n \tag{125}$$

and, taking the derivative,

$$f_{Y_1}(y) = n\left[1 - F_X(y)\right]^{n-1} f_X(y) \tag{126}$$

3.7.2 Asymptotic Extreme Value Distributions

In cases where the cumulative distribution function, $F_X(x)$, is not known (for example the normal or lognormal) explicitly, the exact distributions given above are of questionable value. It turns out that if n is large enough, and the sample is random (ie. independent observations), then the distribution of an extreme value tends towards one of three 'asymptotic' forms. These are explained as follows.

Type I Asymptotic Form

If X has a distribution with an unlimited exponentially decaying tail in the direction of the extreme under consideration, then the distribution of the extreme will tend to the Type I asymptotic form. Examples of such distributions are the normal (in either direction) and the exponential (in the positive direction).

In the case of the maximum, the Type I extreme value distribution has the form

$$\begin{aligned} F_{Y_n}(y) &= \exp\left\{-e^{-\alpha_n(y-u_n)}\right\} \\ f_{Y_n}(y) &= \alpha_n\, e^{-\alpha_n(y-u_n)} \exp\left\{-e^{-\alpha_n(y-u_n)}\right\} \end{aligned} \tag{127}$$

where,

$$\begin{aligned} u_n &= \text{'characteristic' largest value of } X \\ &= F_X^{-1}\left(1-\frac{1}{n}\right) \\ &= \text{mode of } Y_n \\ \alpha_n &= \text{an inverse measure of the variance of } Y_n \\ &= n f_X(u_n) \end{aligned}$$

In particular, u_n is defined as the value that X exceeds with probability $1/n$. It is found by solving $\mathrm{P}\big[X > u_n\big] = 1/n$ for u_n, giving the result shown above. If $F_X^{-1}(p)$ is not known, you will either have to consult the literature or determine this extreme value distribution via simulation.

The mean and variance of the Type I maximum asymptotic distribution are as follows;

$$\begin{aligned} \mathrm{E}\big[Y_n\big] &= u_n + \frac{\gamma}{\alpha_n} \\ \mathrm{Var}\big[Y_n\big] &= \frac{\pi^2}{6\alpha_n^2} \end{aligned} \tag{128}$$

where $\gamma = 0.577216\ldots$ is Euler's number. It always amazes me how often π manages to work its way into results...

The distribution of the minimum value, where the distribution of X is exponentially decaying and unlimited in the direction of the minimum, has the form

$$F_{Y_1}(y) = 1 - \exp\left\{-e^{-\alpha_1(y-u_1)}\right\}$$
$$f_{Y_1}(y) = \alpha_1\, e^{-\alpha_1(y-u_1)} \exp\left\{-e^{-\alpha_1(y-u_1)}\right\} \tag{129}$$

where,

$$\begin{aligned} u_1 &= \text{'characteristic' smallest value of } X \\ &= F_X^{-1}\left(\frac{1}{n}\right) \\ &= \text{mode of } Y_1 \\ \alpha_1 &= \text{an inverse measure of the variance of } Y_1 \\ &= n f_X(u_1) \end{aligned}$$

In particular, u_1 is defined as the value that X has probability $1/n$ of being below. It is found by solving $\mathrm{P}\big[X \le u_1\big] = 1/n$ for u_1. The mean and variance of Y_1 are as follows;

$$\mathrm{E}\big[Y_1\big] = u_1 - \frac{\gamma}{\alpha_1}$$
$$\mathrm{Var}\big[Y_1\big] = \frac{\pi^2}{6\alpha_1^2} \tag{130}$$

Because of the mirror symmetry of the minimum and maximum Type I extreme value distributions, the skewness coefficient of Y_n is 1.1414 whereas the skewness coefficient of Y_1 is -1.1414. That is, the two distributions are mirror images of one another.

Type II Asymptotic Form

If X has a distribution with an unlimited polynomial tail, in the direction of the extreme, then its extreme value will have a type II distribution. Examples of distributions with polynomial tails are the lognormal (in the positive direction) and the Pareto (also in the positive direction) distributions, the latter of which has the form

$$F_X(x) = 1 - \left(\frac{b}{x}\right)^{\alpha}, \quad \text{for } x \ge b \tag{131}$$

If the coefficient b is replaced by $u_n/n^{1/\alpha}$, then we get

$$F_X(x) = 1 - \frac{1}{n}\left(\frac{u_n}{x}\right)^{\alpha}, \quad \text{for } x \ge u_n/n^{1/\alpha} \tag{132}$$

The corresponding extreme value distribution for the maximum, in the limit as $n \to \infty$, is

$$F_{Y_n}(y) = \exp\left\{-\left(\frac{u_n}{y}\right)^{\alpha}\right\}, \quad \text{for } y \geq 0$$
$$f_{Y_n}(y) = \left(\frac{\alpha}{u_n}\right)\left(\frac{u_n}{y}\right)^{\alpha+1} \exp\left\{-\left(\frac{u_n}{y}\right)^{\alpha}\right\} \tag{133}$$

where,

$$\begin{aligned} u_n &= \text{'characteristic' largest value of } X \\ &= F_X^{-1}\left(1 - \frac{1}{n}\right) \\ &= \text{mode of } Y_n \\ \alpha &= \text{shape parameter} \\ &= \text{order of polynomial decay of } F_X(x) \text{ in direction of the extreme} \end{aligned}$$

Note that although the lognormal distribution seems to have an exponentially decaying tail in the direction of the maximum, the distribution is actually a function of the form $a\exp\{-b(\ln x)^2\}$ which has a polynomial decay. Thus, the extreme value distribution of n lognormally distributed random variables follows a type II distribution with

$$\begin{aligned} \alpha &= \frac{\sqrt{2\ln n}}{\sigma_{\ln X}} \\ u_n &= \exp\{u_n'\} \\ u_n' &= \sigma_{\ln X}\sqrt{2\ln n} - \frac{\sigma_{\ln X}\Big(\ln(\ln n) + \ln(4\pi)\Big)}{2\sqrt{2\ln n}} + \mu_{\ln X} \end{aligned} \tag{134}$$

The distribution of the minimum, for an unbounded polynomial decaying tail, can be found as the negative 'reflection' of the maximum, namely as

$$F_{Y_1}(y) = 1 - \exp\left\{-\left(\frac{u_1}{y}\right)^{\alpha}\right\}, \quad y \leq 0;\ u_1 < 0$$
$$f_{Y_1}(y) = -\left(\frac{\alpha}{u_1}\right)\left(\frac{u_1}{y}\right)^{\alpha+1} \exp\left\{-\left(\frac{u_1}{y}\right)^{\alpha}\right\} \tag{135}$$

where,

$$\begin{aligned} u_1 &= \text{'characteristic' smallest value of } X \\ &= F_X^{-1}\left(\frac{1}{n}\right) \\ &= \text{mode of } Y_1 \\ \alpha &= \text{shape parameter} \\ &= \text{order of polynomial decay of } F_X(x) \text{ in direction of the extreme} \end{aligned}$$

Type III Asymptotic Form

If the distribution of X is bounded by a value, say u, in the direction of the extreme, then the asymptotic (as $n \to \infty$) extreme value distribution is the Type III form. For the maximum, this form is

$$F_{Y_n}(y) = \exp\left\{-\left(\frac{u-y}{u-u_n}\right)^{\alpha}\right\}, \qquad \text{for } y \leq u \tag{136}$$

where,

$$\begin{aligned} u_n &= \text{'characteristic' largest value of } X \\ &= F_X^{-1}\left(1-\frac{1}{n}\right) \\ &= \text{mode of } Y_n \\ \alpha &= \text{shape parameter} \\ &= \text{order of polynomial decay of } F_X(x) \text{ in direction of the extreme} \end{aligned}$$

In the case of the minimum, the asymptotic extreme value distribution is

$$F_{Y_1}(y) = 1 - \exp\left\{-\left(\frac{y-u}{u_1-u}\right)^{\alpha}\right\}, \qquad \text{for } y \geq u \tag{137}$$

where,

$$\begin{aligned} u_1 &= \text{'characteristic' smallest value of } X \\ &= F_X^{-1}\left(\frac{1}{n}\right) \\ &= \text{mode of } Y_1 \\ \alpha &= \text{shape parameter} \\ &= \text{order of polynomial decay of } F_X(x) \text{ in direction of the extreme} \end{aligned}$$

and u is now the minimum bound on X. This distribution is also called the Weibull distribution. The shape parameter α is, as mentioned, the order of the polynomial in the direction of the extreme. For example, if X is exponentially distributed, and we are looking at the distribution of the minimum, then $F_X(x)$ has Taylor's series expansion for small x of

$$F_X(x) = 1 - e^{-\lambda x} \simeq 1 - (1 - \lambda x) = \lambda x \tag{138}$$

which has order 1 as $x \to 0$. Thus, for the minimum of an exponential distribution, $\alpha = 1$.

Functions of random variables

Abdul-Hamid Soubra and Emilio Bastidas-Arteaga

University of Nantes – GeM Laboratory, France

In many engineering problems, the uncertainty associated with one random variable needs to be estimated indirectly from the information on uncertainty in another random variable. In most cases, functional relationships (linear or nonlinear) between the response and basic random variables are known; however, in some cases, the exact relationship may not be known explicitly. Since the response variable is a function of other random variables, it will also be random, whether the exact functional relationship between them is known or not. The subject of this chapter is the quantification of the uncertainty in the response variable when it is related to other random variables with a known or unknown relationship.

1 Introduction

This chapter deals with the study of functions of random variable(s). Engineering problems often involve the determination of a relationship between a dependent variable and one or more basic or independent variables. If any one of the independent variables is random, the dependent variable will likewise be random. The probability distribution (as well as its statistical moments) of the dependent variable will be functionally related to and may be derived from those of the basic random variables. As a simple example, the deflection D of a cantilever beam of length L subjected to a concentrated load P (applied at the end of the cantilever) is functionally related to the load P and the modulus of elasticity E of the beam material $[D = (PL^3)/(3EI)]$ in which I is the moment of inertia of the beam cross section. Clearly, we can expect that if P and E are both random variables, with respective $PDFs$, f_P and f_E, the deflection D will also be a random variable with PDF, f_D, that can be derived from the $PDFs$ of P and E. Moreover, the first two statistical moments (i.e. the mean and variance) of D can also be derived as a function of the respective moments of $P$and E.

In this chapter, we shall develop and illustrate the relevant concepts and procedures for determining the $PDF$of the response variable or the statistical moments of this

response variable. Both cases where the functional relationship between the response variable and the independent variables is known or unknown are considered.

2 Exact distributions of functions of random variable(s)

The exact distribution of a function of random variables is considered herein only in case where the response variable is a function of a single random variable. The case where this response variable is a function of several random variables can be found elsewhere (cf. [Hal00], [Ang07] and [Fen08] among others).

2.1 Function of a single random variable

Consider a general case in which the functional relationship between the response variable and the basic random variable is not linear. Assume that the response variable Y is functionally related to X as:

$$Y = g(X) \tag{1}$$

If Y is a monotonically increasing function of X, then

$$P(Y \leq y) = P(X \leq x) \tag{2}$$

Or:

$$F_Y(y) = F_X(x) = F_X[g^{-1}(y)] \tag{3}$$

The value $g^{-1}(y)$ can be evaluated by inverting equation (1). If both sides are differentiated with respect to y, the PDF of Y can be obtained as:

$$f_Y(y) = f_X[g^{-1}(y)]\frac{dg^{-1}(y)}{dy} \tag{4}$$

Thus, if the functional relationship g and the PDF of X are known, the uncertainty in Y in terms of its PDF can be obtained from equation (4).

If Y decreases with X, $dg^{-1}(y)/dy$ is negative (since g^{-1} decreases with Y). Since the PDF of a random variable cannot be negative, its absolute value is of interest. Therefore, to account for both cases, the PDF of Y is written as

$$f_Y(y) = f_X[g^{-1}(y)] \left| \frac{dg^{-1}(y)}{dy} \right| \tag{5}$$

2.2 Example application for an exact distribution of a function of single random variable

Consider a normal variate X with parameters μ and σ; i.e. $N(\mu, \sigma)$ with PDF

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \tag{6}$$

Let $Y = \frac{X-\mu}{\sigma}$. Using equation (5), we determine the PDF of Y as follows: First, we observe that the inverse function is $g^{-1}(y) = \sigma y + \mu$ and $\frac{dg^{-1}}{dy} = \sigma$. Then, according to equation (5), the PDF of Y is

$$f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma} exp\left[\frac{-\frac{1}{2}(\sigma y + \mu - \mu)^2}{\sigma^2}\right] |\sigma| = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} \tag{7}$$

which is the PDF of the standard normal distribution, $N(0,1)$.

3 Moments of functions of random variables

In the previous section, we derived the probability distribution of a function of one random variable. It was shown in literature that the linear function of normal variate remains normal. The product (or quotient) of lognormal variates remains also lognormal (see [Ang07] among others).

In general, the derived probability distributions of the function may be difficult (or even impossible) to derive analytically. Indeed, if the distributions of the X_i's are not known, or if X_1 is normal, X_2 is lognormal, and so on, it is not possible to determine the exact distribution of the response variable Y; however, its mean and variance can still be extracted from the information on the means and variances of the X_i's, giving only limited information on its randomness.

If the functional relationship is linear, then the mean and variance of the response variable can be estimated without any approximation. For nonlinear functional relationships, the mean and variance of the response variable can only be estimated approximately. These are discussed next. Beforehand, remember that the expected value of a function $Z = g(X_1, X_2, \dots, X_n)$ of n random variables, called the mathematical expectation, is given by:

$$E(Z) = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} g(x_1, x_2, \dots, x_n) f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots x_n) dx_1 dx_2 \dots dx_n \tag{8}$$

Below, we shall use equation (8) to derive the moments of linear functions of random variables, as well as the first-order approximate moments of nonlinear functions.

3.1 Mean and variance of a linear function

Consider first the moments of the linear function

$$Y = aX + b \tag{9}$$

According to equation (8), the mean value of Y is:

$$\begin{aligned} E(Y) = E(aX + b) &= \int_{-\infty}^{\infty} (ax + b) f_X(x) dx \\ &= a \int_{-\infty}^{\infty} x f_X(x) dx + b \int_{-\infty}^{\infty} f_X(x) dx = aE(X) + b \end{aligned} \tag{10}$$

whereas the variance is:

$$\begin{aligned} VAR(Y) = E[(Y - \mu_Y)^2] &= E[(aX + b - a\mu_X - b)^2] \\ &= a^2 \int_{-\infty}^{\infty} (x - \mu_X)^2 f_X(x) dx = a^2 Var(X) \end{aligned} \tag{11}$$

For $Y = a_1 X_1 + a_2 X_2$,

where a_1 and a_2 are constants

$$E(Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (a_1 x_1 + a_2 x_2) f_{X_1, X_2}(x_1, x_2) dx_1 dx_2 \tag{12}$$

This equation may be written (in case where X_1 and X_2 are statistically independent) as follows:

$$E(Y) = a_1 \int_{-\infty}^{\infty} x_1 f_{X_1}(x_1) dx_1 + a_2 \int_{-\infty}^{\infty} x_2 f_{X_2}(x_2) dx_2 \tag{13}$$

We can recognize that the last two integrals above are, respectively, $E(X_1)$ and $E(X_2)$; hence, we have for the sum of two random variables

$$E(Y) = a_1 E(X_1) + a_2 E(X_2) \tag{14}$$

The variance of Y (for the general case of correlated random variables) is given by:

$$\begin{aligned} Var(Y) &= E\left[(a_1 X_1 + a_2 X_2) - \left(a_1 \mu_{X_1} + a_2 \mu_{X_2}\right)\right]^2 \\ &= E\left[a_1 \left(X_1 - \mu_{X_1}\right) + a_2 \left(X_2 - \mu_{X_2}\right)\right]^2 \\ &= E\left[{a_1}^2 \left(X_1 - \mu_{X_1}\right)^2 + 2 a_1 a_2 \left(X_1 - \mu_{X_1}\right)\left(X_2 - \mu_{X_2}\right) \right. \\ &\quad \left. + {a_2}^2 \left(X_2 - \mu_{X_2}\right)^2\right] \end{aligned} \tag{15}$$

We may recognize that the expected values of the first and third terms within the brackets are variances of X_1 and X_2, respectively, whereas the middle term is the covariance between X_1 and X_2. Hence, we have:

$$Var(Y) = {a_1}^2 Var(X_1) + {a_2}^2 Var(X_2) + 2 a_1 a_2 COV(X_1, X_2) \tag{16}$$

If the variables X_1 and X_2 are statistically independent, $COV(X_1, X_2) = 0$; thus, equation (16) becomes:

$$Var(Y) = {a_1}^2 Var(X_1) + {a_2}^2 Var(X_2) \tag{17}$$

The results we obtained above can be extended to a general linear function of n random variables, such as

$$Y = \sum_{i=1}^{n} a_i X_i \tag{18}$$

in which the a_i's are constants. For this general case, we obtain the mean and variance of Y as follows:

$$E(Y) = \sum_{i=1}^{n} a_i E(X_i) = \sum_{i=1}^{n} a_i \mu_{X_i} \tag{19}$$

$$Var(Y) = \sum_{i=1}^{n} {a_i}^2 Var(X_i) + \sum_{\substack{i,j=1,..,n \\ i \neq j}} a_i a_j COV(X_i, X_j)$$
$$= \sum_{i=1}^{n} {a_i}^2 {\sigma_{X_i}}^2 + \sum_{\substack{i,j=1,..,n \\ i \neq j}} a_i a_j \rho_{ij} \sigma_{X_i} \sigma_{X_j} \quad (20)$$

in which ρ_{ij} is the correlation coefficient between X_i and X_j.

3.2 Taylor's series and approximate moments of a general function

3.2.1. Function of a single random variable

For a general function of a single random variable X,

$$Y = g(X) \quad (21)$$

the exact moments of Y may be obtained using

$$E(Y) = \int_{-\infty}^{\infty} g(x) f_X(x) dx \quad (22)$$

and

$$Var(Y) = \int_{-\infty}^{\infty} [g(x) - \mu_Y]^2 f_X dx \quad (23)$$

Obviously, the determination of the mean and variance of the function Y with the above relations would require information on the DF $f_X(x)$. In many applications, however, the PDF of X may not be available. In such cases, we seek approximate mean and variance of the function Y as follows:

We may expand the function $g(X)$ in a Taylor series about the mean value of X, that is,

$$g(X) = g(\mu_X) + (X - \mu_X)\frac{dg}{dX} + \frac{1}{2}(X - \mu_X)^2 \frac{d^2 g}{dX^2} + \cdots \quad (24)$$

where the derivatives are evaluated at μ_X.

Now, if we truncate the above series at the linear terms, i.e.,

$$g(X) \cong g(\mu_X) + (X - \mu_X)\frac{dg}{dX} \tag{25}$$

We obtain the first-order approximate mean and variance of Y as

$$E(Y) \cong g(\mu_X) \tag{26}$$

and

$$Var(Y) \cong Var(X)\left(\frac{dg}{dX}\right)^2 \tag{27}$$

We should observe that if the function $g(X)$ is approximately linear (i.e. not highly nonlinear) for the entire range of X, equations (26) and (27) should yield good approximations of the exact mean and variance of $g(X)$. Moreover, when $Var(X)$ is small relative to $g(\mu_X)$, the above approximations should be adequate even when the function $g(X)$ is nonlinear.

3.2.2. Function of multiple random variables

If Y is a function of several random variables,

$$Y = g(X_1, X_2, \dots, X_n) \tag{28}$$

We obtain the approximate mean and variance of Y as follows: Expand the function $g(X_1, X_2, \dots, X_n)$ in a Taylor series about the mean values $(\mu_{X_1}, \mu_{X_2}, \dots, \mu_{X_n})$, yielding

$$\begin{aligned} Y = g(\mu_{X_1}, \mu_{X_2}, \dots, \mu_{X_n}) &+ \sum_{i=1}^{n}(X_i - \mu_{X_i})\frac{\partial g}{\partial X_i} \\ &+ \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(X_i - \mu_{X_i})\left(X_j - \mu_{X_j}\right)\frac{\partial^2 g}{\partial X_i X_j} + \cdots \end{aligned} \tag{29}$$

Where the derivatives are all evaluated at $\mu_{X_1}, \mu_{X_2}, \dots, \mu_{X_n}$. If we truncate the above series at the linear terms, i.e.,

$$Y = g(\mu_{X_1}, \mu_{X_2}, \dots, \mu_{X_n}) + \sum_{i=1}^{n}(X_i - \mu_{X_i})\frac{\partial g}{\partial X_i} \tag{30}$$

We obtain the first-order mean and variance of Y, respectively as follows:

$$E(Y) \cong g(\mu_{X_1}, \mu_{X_2}, \dots, \mu_{X_n}) \tag{31}$$

and

$$Var(Y) \cong \sum_{i=1}^{n} \left(\sigma_{X_i}\right)^2 \left(\frac{\partial g}{\partial X_i}\right)^2 + \sum_{\substack{i,j=1,\ldots,n \\ i \neq j}} \rho_{ij} \sigma_{X_i} \sigma_{X_j} \frac{\partial g}{\partial X_i} \frac{\partial g}{\partial X_j} \tag{32}$$

We observe that if X_i and X_j are uncorrelated (or statically independent) for all i and j, i.e. $\rho_{ij} = 0$, then equation (32) becomes

$$Var(Y) \cong \sum_{i=1}^{n} \left(\sigma_{X_i}\right)^2 \left(\frac{\partial g}{\partial X_i}\right)^2 \tag{33}$$

Equation (33) is a function of both the variances of the independent variables and of the sensitivity coefficients as represented by the partial derivatives.

3.2.3. Example application for the computation of the mean and variance of a general function of several variables

Assume that the random variable Y can be represented by the following relationship:

$$Y = X_1 {X_2}^2 {X_3}^{1/3} \tag{34}$$

Where X_1, X_2 and X_3 are statistically independent random variables with means of 1.0, 1.5, and 0.8, respectively, and corresponding standard deviations of 0.10, 0.20, and 0.15, respectively. Using equations (31) and (33), we find the first-order mean and variance, respectively, to be:

$$E(Y) \approx 1.0 \times 1.5^2 \times (0.8)^{1/3} = 2.0887$$

and

$$\begin{aligned} Var(Y) \approx\ & Var(X_1)\left(\mu_{X_2}{}^2 \times \mu_{X_3}{}^{1/3}\right)^2 + Var(X_2)\left[\mu_{X_1} \times \left(2\mu_{X_2}\right) \times \mu_{X_3}{}^{1/3}\right]^2 \\ & + Var(X_3)\left[\mu_{X_1} \times \mu_{X_2}{}^2 \times \left(\frac{1}{3}\mu_{X_3}{}^{-2/3}\right)\right]^2 \\ & = (0.10)^2\left(1.5^2 \times 0.8^{1/3}\right)^2 + (0.20)^2\left(1.0 \times 2 \times 1.5 \times 0.8^{1/3}\right)^2 \\ & + (0.15)^2\left[1.0 \times 1.5^2 \times (1/3) \times 0.8^{-2/3}\right]^2 \\ & = (0.10)^2(2.09)^2 + (0.20)^2(2.78)^2 + (0.15)^2(0.87)^2 \\ & = 0.04363 + 0.31024 + 0.01704 = 0.37091 \end{aligned}$$

and

$$\sigma_Y = 0.609$$

3.3 Mean and variance of an analytically-unknown functional relationship

In many cases, the exact form of g in equation (28) may not be known. In fact, the exact functional relationship is known in algorithmic form but not in any exact functional form. The implication is that the partial derivatives of the function with respect to the random variables cannot be calculated to approximate the first-order mean and the first-order variance of the response variable, as discussed before.

In this case, the approximate (first-order) mean value of the response, represented by Y in equation (28), can be obtained by using the mean values of all the parameters in the problem, the same as in equation (31). Evaluating the variance of Y will be more involved since the functional form of g is unknown, and its partial derivatives with respect to the i^{th} random variable in equation (28) cannot be evaluated. The task is to calculate the variance of Y without information on the analytical partial derivatives. The Taylor series finite difference estimation procedure can be used to numerically evaluate the variance of Y, as discussed below:

To evaluate the variance, one needs to compute, for each random variable, the two following (intermediate) response variables:

$$Y_i^+ = g\left[\mu_{X_1}, \mu_{X_2}, \ldots, \left(\mu_{X_i} + \sigma_{X_i}\right), .., \mu_{X_n}\right] \tag{35}$$

and

$$Y_i^- = g\left[\mu_{X_1}, \mu_{X_2}, \ldots, \left(\mu_{X_i} - \sigma_{X_i}\right), .., \mu_{X_n}\right] \tag{36}$$

In simple terms, equation (35) states that the response variable Y_i^+ is calculated considering the mean of all the random variables except the i^{th} one, which is considered to be the mean plus one standard deviation value. Equation (36) indicates the same thing, except that for the i^{th} random variable, the mean minus one standard deviation value needs to be considered. Then, using the central difference approximation, we can show that

$$E_i = \frac{\partial g}{\partial X_i} = \frac{Y_i^+ - Y_i^-}{2\sigma_{X_i}} \tag{37}$$

Considering all the random variables, the first-order variance of Y is computed as

$$Var(Y) \approx \sum_{i=1}^{n} \left(\frac{Y_i^+ - Y_i^-}{2\sigma_{X_i}}\right)^2 \times Var(X_i) \approx \sum_{i=1}^{n} \left(\frac{Y_i^+ - Y_i^-}{2}\right)^2 \tag{38}$$

Thus, when the functional relationship among the random variables is not known explicitly, the mean and variance of the response variable can be approximated by

using equations (31) and (38). This requires the computation of the response variable several times. If there are n random variables present in a problem, the required total number of computations of the response variable is $(1 + 2n)$.

4. Conclusion

The probabilistic characteristics of a function of random variable(s) may be derived from those of the independent variates. These include, in particular, the probability distribution and the first two statistical moments (mean and variance) of the function. It was shown that for a function of a single random variable, the PDF of the function can be readily obtained analytically. However, it was shown in the literature that the derivation of the distribution of a function of multiple variables can be complicated mathematically, especially for nonlinear functions (see [Ang07] among others). Therefore, even though the required distribution of a function may theoretically be derived, it is often impractical to apply, except for special cases, such as a linear function of independent Gaussian variates or the strictly product/quotient of independent lognormal variates. In this light, it is often necessary, in many applications, to describe the probabilistic characteristics of a function approximately in terms only of its mean and variance. The mean and variance of linear functions can be estimated without any approximation; however, for a general nonlinear function, we must often resort to first-order (or second-order) approximations. In case of analytically-unknown functions, one may use the finite difference method for the (approximate) computation of the statistical moments. Finally, it should be mentioned that when the probability distribution of a general function is required, we may need to resort to Monte Carlo simulations or other numerical methods.

References

[Ang07] A. Ang and W. Tang. Probability concepts in engineering, Emphasis on applications to civil and environmental engineering, John Wiley & Sons, 406 pages, 2007.

[Fen08] G.A. Fenton and D.V. Griffiths. Risk assessment in geotechnical engineering, John Wiley & Sons, 461 pages, 2008.

[Hal00] A. Haldar and S. Mahadevan. Probability, Reliability and Statistical Methods in Engineering desing, Jonh Wiley & Sons, Inc. 2000.

Reliability Analysis Methods

Emilio Bastidas-Arteaga and Abdul-Hamid Soubra

University of Nantes – GeM Laboratory, France

There are numerous sources of uncertainties that should be considered in engineering design. Reliability analysis methods provide a framework to account for these uncertainties in a rational manner. This chapter presents the First Order Reliability Method (FORM) and the Second Order Reliability Method (SORM). These methods are illustrated by academic examples.

1. Introduction

Engineering design aims at providing minimum levels of serviceability and safety during the structural lifetime. This is a difficult task because there are important sources of uncertainty that could lead to over- or under-design solutions. For example, there are uncertainties related to environmental exposure, loading, material properties, engineering models, etc. Reliability analysis methods offer the theoretical framework for considering uncertainties in a comprehensive decision scheme. The main goal of reliability analysis methods is to evaluate the ability of systems or components to remain safe and operational during their lifecycle.

The main objective of this chapter is to present and illustrate various reliability analysis methods that can be used for engineering or research purposes. The chapter starts with a description of fundamental concepts for reliability analysis. After, we present the First Order Reliability Method (FORM) and the Second Order Reliability Method (SORM). These methods are illustrated by academic examples.

2. Basic concepts for reliability analysis

Reliability methods have been established to take into account, in a rigorous manner, the uncertainties involved in the analysis of an engineering problem. The *failure probability* and the *reliability index* are used to quantify risks and therefore evaluate the consequences of failure. In this approach, the governing parameters of the prob-

lem are modeled as random variables. Random variables can be grouped in a random vector $\mathbf{X}$ where $f_{\mathbf{X}}(\mathbf{x})$ is the joint probability density function (PDF).

For reliability analysis, the space D of random variables may be divided into the *failure* and the *safety regions*. The failure region D_f is defined by $D_f = \{\mathbf{X} \mid g(\mathbf{X}) \leq 0\}$ and the *safety region*, D_s, by $D_s = \{\mathbf{X} \mid g(\mathbf{X}) > 0\}$ where $g(\mathbf{X})$ represents the performance function. Notice that $g(\mathbf{X})=0$ is the boundary between failure and safety regions and it is called the *limit state surface*.

In the simplest case, the performance function $g(\mathbf{X})$ is expressed as the difference between the resistance $R(\mathbf{X})$ and the demand or solicitation on the system $S(\mathbf{X})$ – i.e., $g(\mathbf{X}) = R(\mathbf{X}) - S(\mathbf{X})$. In reliability engineering analysis, $g(\mathbf{X})$ is usually expressed in terms of displacement, strain, stress, etc. The performance functions can be related to the following structural conditions:

1. *Serviceability limit state*: under this condition, 'failure' is related to a serviceability loss that does not imply a significant decay of structural safety. For example, if the reliability analysis of a given structural component focuses on a maximum displacement v_{max}, the performance function can write:

$$g(X) = v_{max} - v(\boldsymbol{X}) \tag{1}$$

   where v_{max} could be fixed by standards or particular serviceability constraints and $v(\mathbf{X})$ is the displacement of the point of interest that depends on $\mathbf{X}$ random variables (material strength, geometry, load, etc.). In the case of failure, $v(\mathbf{X}) > v_{max}$, but the structural component is still considered safe.

2. *Ultimate limit state*: this condition describes the state at which structural safety is highly affected and may lead to total failure or collapse. For instance, if the reliability analysis focuses on the bending moment of a beam, the performance function is:

$$g(X) = M_r(\boldsymbol{X}) - M_s \tag{2}$$

   where $M_r(\mathbf{X})$ is the resistant bending moment of the beam that depends on $\mathbf{X}$ random variables (material strength, sectional geometry, etc.), and M_s is the soliciting bending moment. Notice that although M_s is assumed to be deterministic in eq. (2), this variable may also be considered as a random variable. In the case of failure, $M_s > M_r(\mathbf{X})$, leading to the collapse of the beam.

By accounting for these definitions, the failure probability, P_f, is determined by:

$$P_f = \mathrm{P}[g(\mathbf{X}) \leq 0] = \int_{g(\mathbf{X})\leq 0} f_{\mathbf{X}}(\mathbf{x}) dx_1 \ldots dx_n \tag{3}$$

Notice that the limit state function can be a linear or a nonlinear function of the basic variables. FORM can be used to evaluate eq. (3) when the limit state function is a linear function of uncorrelated normal variables or when the nonlinear limit state function is represented by a first-order (linear) approximation with equivalent normal variables. SORM estimates the probability of failure by approximating the nonlinear limit state function (including a linear limit state function with correlated non-normal variables) by a second-order representation.

3. First Order Reliability Methods (FORM)

The First Order Reliability Method (FORM) makes use of the first and second moments of the random variables. This method includes two approaches [Hal00]. These are First-Order Second-Moment (FOSM) and Advanced First-Order Second-Moment (AFOSM) approaches. In FOSM, the information on the distribution of random variables is ignored; however, in AFOSM, the distributional information is appropriately used.

3.1. First Order Second Moment (FOSM)

The First Order Second Moment (FOSM) method makes use of only second-moment statistics (i.e. mean and standard deviation) of the random variables and it requires a linearized form of the performance function at the mean values of the random variables. A first-order Taylor series approximation is used to linearize the performance function at the mean values of the random variables.

Cornell [Cor69] proposed the original FOSM formulation. Let us consider an elementary reliability case where a weight is hung by a cable. The load-carrying capacity or the resistance of the cable is R and the load effect is S. The resistance and the load will be modeled as independent Gaussian random variables with $N(\mu_R, \sigma_R)$ and $N(\mu_S, \sigma_S)$, respectively. In this case, the failure probability P_f is related to the failure event $R < S$, and is computed as:

$$P_f = P(R < S) = P(R - S < 0) \tag{4}$$

A new random variable Z (called performance function) can be introduced:

$$Z = R - S \tag{5}$$

The performance function Z is characterized by a mean $\mu_Z = \mu_R - \mu_S$ and a standard deviation $\sigma_Z^2 = \sigma_R^2 + \sigma_S^2$. Since R and S are Gaussian, it can be demonstrated that Z also follows a Gaussian distribution. Figure 1 presents the PDF of Z. It is observed that the failure probability is related to the event $P(Z < 0)$. Consequently, P_f could be estimated directly from:

$$P_f = \Phi[(0 - \mu_Z)/\sigma_Z] = \Phi[-\mu_Z/\sigma_Z] = \Phi[-\beta] \tag{6}$$

where Φ[.] represents the standard normal cumulative distribution function (CDF) and $\beta = \mu_Z / \sigma_Z$ is the 'reliability index' that is also used to quantify risks of failure.

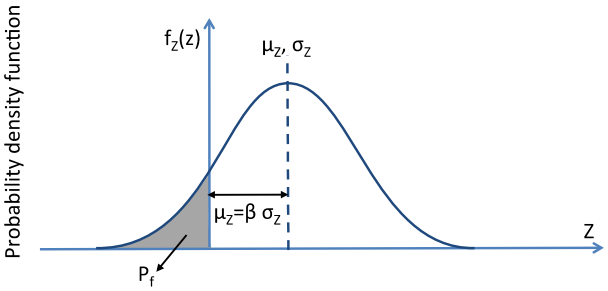


Figure 1: Probability density function of Z

As may be easily seen from Figure 1, the reliability index computed by FOSM represents the number of standard deviations that separate the mean value of the performance function from the limit state surface $Z = 0$.

For lognormal random variables, an alternative formulation to eq. (6) could be derived as follows: Assume that R and S are statically independent lognormal variables, that is, $LN(\lambda_R, \xi_R)$ and $LN(\lambda_S, \xi_S)$. In this case, another random variable Y can be introduced as

$$Y = R / S \tag{7}$$

or

$$\ln Y = Z = \ln R - \ln S \tag{8}$$

The failure event can be defined as $Y < 1.0$ or $Z < 0$. Since R and S are lognormal, $\ln R$ and $\ln S$ are normal; therefore, $\ln Y$ or Z is a normal random variable with mean $\lambda_R - \lambda_S$ and standard deviation $\sqrt{\xi_R^{\,2} + \xi_S^{\,2}}$. The probability of failure can be defined (similar to eq. (6)) by

$$P_f = \Phi\left[\frac{0-(\lambda_R-\lambda_S)}{\sqrt{\xi_R^{\,2}+\xi_S^{\,2}}}\right] = \Phi\left[-\frac{\mu_Z}{\sigma_Z}\right] = \Phi[-\beta] \tag{9}$$

In the general case where the performance function Z is a function of a vector of n random variables, i.e.,

$$Z = g(\mathbf{X}) = g(X_1, X_2, \ldots, X_n) \tag{10}$$

the Taylor series expansion about the mean value gives:

$$Z = g(\mu_{\mathbf{X}}) + \sum_{i=1}^{n} \frac{\partial g}{\partial X_i}(X_i - \mu_{X_i}) + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\partial^2 g}{\partial X_i \partial X_j}(X_i - \mu_{X_i})(X_j - \mu_{X_j}) + \ldots \quad (11)$$

where the derivatives are evaluated at the mean values of the random variables (X_1, X_2,…, X_n), and μ_{X_i} is the mean value of X_i. By truncating eq. (11) at the linear terms, it is possible to obtain first-order approximations of the mean and variance of Z as follows:

$$\mu_Z \approx g(\mu_{X_1}, \mu_{X_2}, ..., \mu_{X_n}) \quad (12)$$

and

$$\sigma_Z^2 \approx \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\partial g}{\partial X_i}\frac{\partial g}{\partial X_j}\mathrm{Cov}(X_i, X_j) \quad (13)$$

where $\mathrm{Cov}(X_i, X_j)$ is the covariance of X_i and X_j. For uncorrelated random variables, the variance becomes:

$$\sigma_Z^2 \approx \sum_{i=1}^{n}\left(\frac{\partial g}{\partial X_i}\right)^2 \mathrm{Var}(X_i) \quad (14)$$

Consequently, by estimating μ_Z and σ_Z from eqs. (12) and (13), respectively, the reliability index can be computed as $\beta = \mu_Z / \sigma_Z$.

The exact failure probability could be derived from the reliability index only in few cases:

1. if all the X_i*'s* are statically independent normal variables and if Z is a linear function of the X_i, then, Z is normal and the probability of failure is given by eq. (6);
2. if all the X_i*'s* are statically independent lognormal variables and if g(**X**) is a multiplicative function of the X_i*'s*, then $Z = \ln g(\mathbf{X})$ is normal and the probability of failure is also given by $P_f = \Phi(-\beta)$.

To conclude, in most cases it is not likely that all the variables are statically independent normal or lognormal. Nor is it likely that the performance function is a simple additive or multiplicative function of these variables. In such cases, the reliability index cannot be directly related to the probability of failure; nevertheless, the equation $P_f = \Phi(-\beta)$ does provide a rough idea of the level of risk. Notice finally that FOSM approach has the following shortcomings:

- the information about the distribution of the independent variables is not considered,
- the truncation errors may be significant if $g(.)$ is non-linear,

- the assessment of reliability index varies when the limit state function is expressed in different but mechanically equivalent formulations – e.g., ($R - S = 0$) or ($R / S = 1$).

Example 1 *Reliability index of a steel beam using FOSM*

Let us consider a steel beam subjected to a deterministic bending moment of M_a =130 kNm. The yield stress F_y and the plastic modulus Z_p of the beam are considered as random variables following normal distributions with the following parameters:

$$\mu_{F_y} = 250 \text{ MPa}, \sigma_{F_y} = 25 \text{ MPa}$$
$$\mu_{Z_p} = 9\times10^{-4} \text{ m}^3, \sigma_{Z_p} = 4.5\times10^{-5} \text{ m}^3$$

Question 1: Write two different performance functions by considering strength and stress formulations.

For the strength formulation, the resistance of the beam is a random variable defined as $R = F_y Z_p$ and the solicitation S is deterministic where $S = M_a$. Then, the performance function becomes

$$g(F_y, Z_p) = R - S = F_y Z_p - M_a \tag{15}$$

For the stress formulation, the yield stress becomes the resistance of the performance function, i.e. $R = F_y$ and the solicitation is computed as $S = M_a/Z$. In this case, both R and S are random variables and the performance function becomes

$$g(F_y, Z_p) = R - S = F_y - \frac{M_a}{Z_p} \tag{16}$$

Question 2: Estimate the reliability index using the strength formulation.

We assume that the random variables are independent. We will first estimate μ_R and σ_R by using eqs. (12) and (14) respectively:

$$\mu_R \approx \mu_{F_y}\mu_{Z_p} = (250)(9\times10^{-4}) = 225 \text{kN}\cdot\text{m}$$

$$\sigma_R \approx \left[\text{Var}(F_y)\left(\frac{\partial R}{\partial F_y}\right)^2 + \text{Var}(Z_p)\left(\frac{\partial R}{\partial Z_p}\right)^2\right]^{\frac{1}{2}} = \left[25^2\left(9\times10^{-4}\right)^2 + \left(4.5\times10^{-5}\right)^2 250^2\right]^{\frac{1}{2}}$$
$$= 25.16 \text{ kN}\cdot\text{m}$$

The sollicitation is deterministic, then, $\mu_S = M_a$ and $\sigma_S = 0$. Consequently, the reliability index is given by :

$$\beta = \frac{\mu_R - \mu_S}{\sqrt{\sigma_R^2 + \sigma_S^2}} = \frac{225 - 130}{\sqrt{25.15^2 + 0^2}} = 3.77$$

Question 3: Estimate the reliability index using the stress formulation.

According to the performance function given by eq. (16), we obtain directly the mean and the standard deviation of the resistance: $\mu_R = \mu_{F_y}$ and $\sigma_R = \sigma_{F_y}$. For the solicitation, μ_S and σ_S are computed from eqs. (12) and (14) respectively:

$$\mu_S \approx M_a / \mu_{Z_p} = 130 / (9 \times 10^{-4}) = 144.4 \text{ MPa}$$

$$\sigma_S \approx \left[\text{Var}(Z_p) \left(-\frac{M_a}{\mu_{Z_p}^2} \right)^2 \right]^{\frac{1}{2}} = \sigma_{Z_p} \frac{M_a}{\mu_{Z_p}^2} = 4.5 \times 10^{-5} \frac{130}{\left(9 \times 10^{-4}\right)^2} = 7.22 \text{ MPa}$$

And the reliability index becomes

$$\beta = \frac{\mu_R - \mu_S}{\sqrt{\sigma_R^2 + \sigma_S^2}} = \frac{250 - 144.4}{\sqrt{25^2 + 7.22^2}} = 4.06$$

By comparing the reliability indexes, it is noted that the result depends on the formulation of the performance function. This lack of invariance motivated the development of other reliability methods such as that presented in the following sections.

3.2. Advanced First Order Second Moment (AFOSM)

AFOSM is also called 'Hasofer-Lind' method. In this method, the assessment of the reliability index is mainly based on the transformation/reduction of the problem to a standardized coordinate system. Thus, a random variable X_i is reduced as:

$$X_i' = (X_i - \mu_{X_1}) / \sigma_{X_1} \quad (i = 1,2,...,n) \tag{17}$$

where X_i' is a random variable with zero mean and unit standard deviation. Thus, Eq. (17) is used to transform the original limit state surface $g(\mathbf{X}) = 0$ into a reduced limit state surface $g(\mathbf{X'}) = 0$. Consequently, $\mathbf{X}$ denotes 'original coordinate system' and $\mathbf{X'}$ describes the 'transformed or reduced coordinate system'. In the standardized coordinate system, the Hasofer-Lind reliability index β_{HL} corresponds to the minimum distance from the origin of the axes (in the reduced coordinates system) to the limit state surface:

$$\beta_{HL} = \sqrt{(\mathbf{x}'^*)^T (\mathbf{x}'^*)} \tag{18}$$

The minimum distance point on the limit state surface is called the 'design point'. It is denoted by vector **x*** in the original coordinate system and by vector **x'*** in the reduced coordinate system. These vectors represent the values of all the random variables, *i.e.* $X_1, X_2, \ldots, X_n$ at the design point corresponding to the coordinate system being used.

Figure 2 illustrates the reduction of the random variables R and S for a linear performance function such as that described by eq. (5). According to eq. (17), R and S can be reduced as:

$$R' = (R - \mu_R)/\sigma_R \quad \text{and} \quad S' = (S - \mu_S)/\sigma_S \tag{19}$$

The substitution of R' and S' into the limit state surface (Z=0) gives the new limit state surface in the reduced coordinate system (Figure 2b):

$$Z = \sigma_R R' - \sigma_S S' + \mu_R - \mu_S = 0 \tag{20}$$

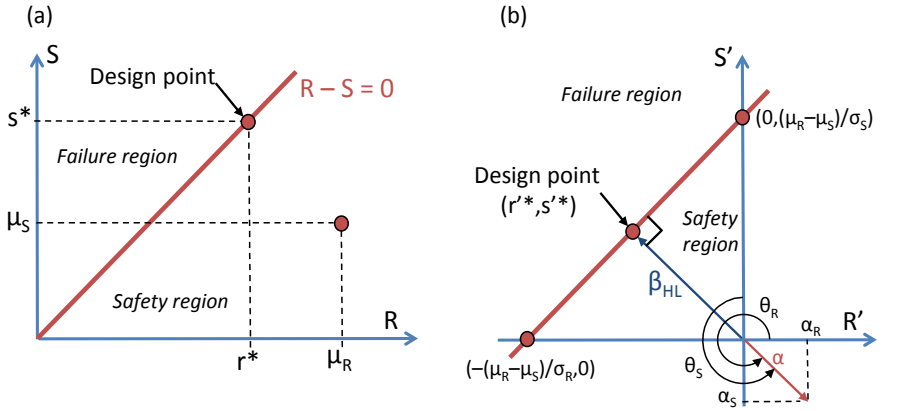


Figure 2: Reduction of coordinates: (a) original coordinates, (b) reduced coordinates

The reliability of the problem is estimated by using eq. (18). It corresponds to the minimum distance between the limit state surface and the origin (in the reduced coordinates system). By using simple trigonometry, this distance (reliability index) can be estimated as:

$$\beta_{HL} = \frac{\mu_R - \mu_S}{\sqrt{\sigma_R^2 + \sigma_S^2}} \tag{21}$$

It should be emphasized here that in the present case of a linear limit state surface, the Hasofer-Lind reliability index corresponds exactly to the reliability index com-

puted from FOSM if both R and S are normal variables. However, this is not the case for other limit state surfaces or random variable distributions.

From Figure (2b), it is apparent that if the limit state line is closer to the origin in the reduced coordinate system, the failure region is larger, and if it is farther away from the origin, the failure region is smaller. Thus, the position of the limit state surface relative to the origin in the reduced coordinate system is a measure of the reliability of the system. Notice that the Hasofer-Lind reliability index can be used to calculate the failure probability as $P_f = \Phi(-\beta_{HL})$. This is the integral of the standard normal density function along the ray joining the origin and **x'***. It is obvious that the nearer **x'*** is to the origin, the larger is also the most probable failure point. The point of minimum distance from the origin to the limit state surface, **x'***, represents the worst combination of the stochastic variables and is appropriately named the 'design point' or the 'most probable point MPP' of failure.

Finally, it should be noted that the Hasofer-Lind reliability index is invariant, because regardless of the form in which the limit state equation is written [e.g., (R–S=0) or (R/S=1)], its geometric shape and the distance from the origin remain constant.

For the general case of a non-linear limit state surface, the assessment of the minimum distance can be written as an optimization problem:

$$\begin{aligned} &\text{minimize } D = \sqrt{\mathbf{x}'^T \mathbf{x}'} \\ &\text{subject to } g(\mathbf{x}') = 0 \end{aligned} \tag{22}$$

By using Lagrange's multipliers, the minimum distance (for n random variables) could be estimated as:

$$\beta_{HL} = -\frac{\sum_{i=1}^{n} x_i'^* \left(\frac{\partial g}{\partial X_i'}\right)^*}{\sqrt{\sum_{i=1}^{n} \left(\frac{\partial g}{\partial X_i'}\right)^{2^*}}} \tag{23}$$

where $(\partial g/\partial X'_i)^*$ is the i^{th} partial derivative evaluated at the design point ($x_1'^*$, $x_2'^*$,…, $x_n'^*$). The design point in the reduced coordinates is:

$$x_i'^* = -\alpha_i \beta_{HL} \quad (i = 1,2,...,n) \tag{24}$$

where α_i are the direction cosines along the coordinate axes X'_i. They are given by:

$$\alpha_i = \frac{\left(\frac{\partial g}{\partial X_i'}\right)^*}{\sqrt{\sum_{i=1}^{n}\left(\frac{\partial g}{\partial X_i'}\right)^{2*}}} \tag{25}$$

By using eq. (17), the design point (in the original space) is given by:

$$x_i^* = \mu_{X_i} - \alpha_i \sigma_{X_i} \beta_{HL} \tag{26}$$

An algorithm was formulated by [Rac76] to compute β_{HL} and $x_i'^*$ as follows:

Step 1: Define the appropriate performance function $g(\mathbf{X})$.

Step 2: Assume initial values for the design point $\mathbf{x}^*_i$. The initial design point is usually assumed to be at the mean value of the n –1 random variables. For the last random variable, its value is obtained from the performance function to ensure that the design point is located on the limit state surface $g(\mathbf{X}) = 0$.

Step 3: Obtain the design point in the reduced space $\mathbf{x}'^* = [x_1'^*, x_2'^*, \ldots, x_n'^*]$ as

$$x_i'^* = (x_i^* - \mu_{X_i}) / \sigma_{X_i} \quad (i = 1,2,...,n) \tag{27}$$

Step 4: Estimate the partial derivatives of the performance function $(\partial g/\partial X'_i)^*$ with respect to the variables in the reduced space and evaluate them at $x_i'^*$. These derivatives can be estimated from the performance function in the original space by using the chain rule of differentiation

$$\frac{\partial g}{\partial X_i'} = \frac{\partial g}{\partial X_i}\frac{\partial X_i}{\partial X_i'} = \frac{\partial g}{\partial X_i}\sigma_{X_i} \tag{28}$$

Define the column vector $\mathbf{A}$ such that

$$A_i = \left(\frac{\partial g}{\partial X_i'}\right)^* \tag{29}$$

Step 5: Compute the reliability index as

$$\beta_{HL} = -\frac{\mathbf{A}^T \mathbf{x}'^*}{\sqrt{\mathbf{A}^T \mathbf{A}}} \tag{30}$$

Step 6: Determine a vector of directional cosines as

$$\boldsymbol{\alpha} = \frac{\mathbf{A}}{\sqrt{\mathbf{A}^T \mathbf{A}}} \tag{31}$$

Step 7: Obtain the new design point $x_i'^*$ for the $n-1$ random variables.

Step 8: Determine the coordinates of the new design point in the original space for the $n-1$ random variables considered in Step 7 as

$$x_i^* = \mu_{X_i} + x_i'^* \sigma_{X_i} \quad (i = 1,2,...,n) \tag{32}$$

Step 9: Determine the value of the last random variable in the original space such that the estimated point belongs to the limit state surface $g(\mathbf{X}) = 0$.

Step 10: Repeat Steps 3 to 9 until convergence of β_{HL}.

Finally, notice that the Hasofer-Lind reliability index β_{HL} can be used to estimate a first-order approximation of the failure probability as $P_f \approx \Phi(-\beta_{HL})$.

Example 2 *Assessment of the reliability index by using FORM (Adapted from [San10])*

Let us suppose that the performance function of a problem is defined by

$$g(X_1, X_2, X_3) = 6.2\, X_1 - X_2\, X_3^2$$

where the random variables X_1, X_2, X_3 follow normal distributions with means $\mu_{X1} = 20$, $\mu_{X2} = 5$, and $\mu_{X3} = 4$; and standard deviations $\sigma_{X1} = 3.5$, $\sigma_{X2} = 0.8$, and $\sigma_{X3} = 0.4$.

Question 1: Estimate the reliability index by using FORM.

Once the performance function is defined (Step 1), we define the coordinates of the design point in the original space (Step 2):

$$\mathbf{x}^* = \left[x_1^* = \mu_{X_1}, x_2^* = \mu_{X_2}, x_3^* = \sqrt{\frac{6.2 x_1^*}{x_2^*}} \right] = [20,5,4.979]$$

The Step 3 consists of obtaining the design points in the reduced space:

$$x_1'^{*} = (x_1^{*} - \mu_{X_1}) / \sigma_{X_1} = 0$$
$$x_2'^{*} = (x_2^{*} - \mu_{X_2}) / \sigma_{X_2} = 0$$
$$x_3'^{*} = (4.979 - \mu_{X_3}) / \sigma_{X_3} = 2.448$$

Then, $(\mathbf{x'}^{*})^{T} = [0, 0, 2.448]$. In the Step 4, we estimate the vector containing the derivatives of the performance function evaluated at the design point (eqs. 28-29):

$$A_1 = \left(\frac{\partial g}{\partial X_1'}\right)^{*} = \left(\frac{\partial g}{\partial X_1}\right)^{*} \sigma_{X_1} = (6.2)(3.5) = 21.7$$
$$A_2 = \left(\frac{\partial g}{\partial X_2'}\right)^{*} = \left(\frac{\partial g}{\partial X_2}\right)^{*} \sigma_{X_2} = -(x_3^{*})^2 \sigma_{X_2} = -19.84$$
$$A_3 = \left(\frac{\partial g}{\partial X_3'}\right)^{*} = \left(\frac{\partial g}{\partial X_3}\right)^{*} \sigma_{X_3} = -2x_2^{*}x_3^{*}\sigma_{X_3} = -19.92$$

Therefore, $\mathbf{A}^{T} = [21.7, -19.84, -19.92]$. The first estimate of the reliability index becomes (Step 5):

$$\beta_{HL} = -\frac{\mathbf{A}^{T}\mathbf{x}'^{*}}{\sqrt{\mathbf{A}^{T}\mathbf{A}}} = 1.374$$

We estimate the vector of directional cosines in Step 6:

$$\boldsymbol{\alpha} = \frac{\mathbf{A}}{\sqrt{\mathbf{A}^{T}\mathbf{A}}} = \begin{bmatrix} 0.611 \\ -0.559 \\ -0.561 \end{bmatrix}$$

The new design point in the reduced space is estimated for the $n - 1$ random variables (Step 7):

$$x_1'^{*} = -\alpha_1\beta_{HL} = (-0.611)(1.374) = -0.840$$
$$x_2'^{*} = -\alpha_2\beta_{HL} = (0.559)(1.374) = 0.768$$

Consequently, the corresponding values in the original space are (Step 8):

$$x_1^{*} = \mu_{X_1} + x_1'^{*}\sigma_{X_1} = 20 + (-0.840)3.5 = 17.06$$
$$x_2^{*} = \mu_{X_2} + x_2'^{*}\sigma_{X_2} = 5 + (0.768)0.8 = 5.614$$

We estimate afterwards the value of x_3^* such that it belongs to the limit state function (Step 9):

$$x_3^* = \sqrt{\frac{6.2x_1^*}{x_2^*}} = 4.34$$

The algorithm is repeated until convergence of the reliability index (Step 10). Table 1 presents the results of various iterations. Convergence is reached after three iterations for this example. We found a reliability index $\beta_{HL} = 1.413$ that corresponds to a failure probability $P_f \approx 0.079$ computed based on a first-order approximation.

Table 1: Summary of the iterative process to estimate β_{HL}

Variable	Iteration Number		
	1	2	3
x_1^*	20	17.06	16.732
x_2^*	5	5.614	5.519
x_3^*	4.98	4.34	4.335
β_{HL}	1.374	1.413	1.413
x_1^*	17.06	16.732	16.709
x_2^*	5.614	5.519	5.521
x_3^*	4.34	4.335	4.332

3.2.1. Extension of AFOSM to the case of non-normal random variables

To extend the Hasofer-Lind method to the case of non-normal random variables, Rackwitz and Fiessler [Rac78] proposed to transform each non-normal random variable into an equivalent normal random variable with a mean $\mu_{X_i}^e$ and standard deviation $\sigma_{X_i}^e$. This transformation allows estimating a solution in the reduced space by using the procedure explained in the previous paragraphs. The equivalent parameters evaluated at the design point x_i^* are given by:

$$\mu_{X_i}^e = x_i^* - \Phi^{-1}\left[F_{X_i}\left(x_i^*\right)\right]\sigma_{X_i}^e \tag{33}$$

$$\sigma_{X_i}^e = \frac{\phi\left[\Phi^{-1}\left[F_{X_i}\left(x_i^*\right)\right]\right]}{f_{X_i}\left(x_i^*\right)} \tag{34}$$

where $\Phi[.]$ and $\phi[.]$ are the CDF and the PDF of the standard variate, respectively, and $F_{X_i}(.)$ and $f_{X_i}(.)$ are the CDF and PDF of the original non-normal random vari-

able (some useful Matlab commandes may be found in Appendix). Notice that eqs. (33-34) are derived by equating the cumulative distribution functions and the probability density functions of the actual variables and the equivalent normal variables at the design point on the limit state surface.

Since the equivalent parameters are evaluated at the design point, each iteration should include the assessment of the equivalent parameters. The before presented algorithm is thus modified as follows:

- Steps 1 and 2 remain similar.
- Step 3 should include the assessment of the equivalent parameters $\mu_{X_i}^e$ and $\sigma_{X_i}^e$ at the design point for each non-normal random variable. These equivalent parameters will be used to determine the design point in the reduced space as follows

$$x_i'^* = (x_i^* - \mu_{X_i}^e) / \sigma_{X_i}^e \tag{35}$$

- Steps 4 to 7 remain exactly similar.
- In Step 8, the assessment of the coordinates in the original space becomes

$$x_i^* = \mu_{X_i}^e + x_i'^* \sigma_{X_i}^e \tag{36}$$

- Finally, Steps 9 and 10 remain similar.

3.3. Ellipsoid Approach

The Hasofer-Lind reliability index can be formulated in a matrix form:

$$\beta_{HL} = \min_{g(\mathbf{X})=0} \sqrt{(\mathbf{X}-\boldsymbol{\mu})^T \mathbf{C}^{-1} (\mathbf{X}-\boldsymbol{\mu})} \tag{37}$$

where $\mathbf{X}$ is a vector containing the n random variables, $\boldsymbol{\mu}$ is a vector containning their mean values, and $\mathbf{C}$ is the covariance matrix.

An intuitive interpretation of the reliability index was suggested in Low and Tang (cf. [Low97] and [Low04]), where the concept of an expanding ellipse (Figure 3) led to a simple method for computing the Hasofer-Lind reliability index in the original space of the random variables using an optimization tool available in most spreadsheet software packages.

When there are only two uncorrelated non-normal random variables X_1 and X_2, these variables span a two-dimensional random space, with an equivalent one-sigma dispersion ellipse (corresponding to β_{HL}=1 in Eq. (37) without the min.), centered at the equivalent mean values $\mu_1{}^N$ and $\mu_2{}^N$ and whose axes are parallel to the coordinate axes of the original space. For correlated variables, a tilted ellipse is obtained.

Low and Tang (cf. [Low97] and [Low04]) reported that the Hasofer-Lind reliability index β_{HL} may be regarded as the codirectional axis ratio of the smallest ellipse (which is either an expansion or a contraction of the 1−σ ellipse) that just touches the limit state surface to the 1−σ dispersion ellipse. They also stated that finding the smallest ellipse that is tangent to the limit state surface is equivalent to finding the most probable failure point.

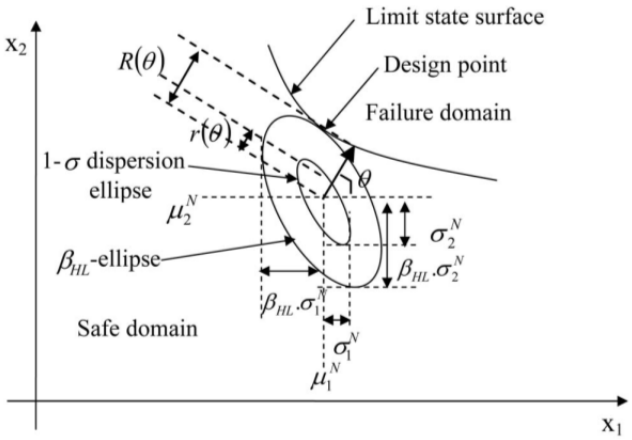


Figure 3: Ellipsoid approach for the computation of the Hasofer-Lind reliability index

3.3.1. Ellipsoid Approach via Spreadsheet

Low and Tang (cf. [Low97] and [Low04]) showed that the minimization of the Hasofer-Lind reliability index can be efficiently carried out in the Excel spreadsheet environment. The spreadsheet approach is simple and easy to understand because it works in the original space of random variables and does not require the additional step of transforming **X** to **X***'* where **X***'* is a transformed vector of the random variables in the uncorrelated Gaussian space. Notice however that the optimization in original space is not preferred from a computational perspective. This is because the optimization in a standardized space is mathematically more desirable in nonlinear optimization. For example, when minimizing the quadratic form of Eq. (37) in the original space; in some cases, the correct solution is obtained only when the solver option "use automatic scaling" is activated. As an alternative, Cholesky factorization of the covariance matrix can be used.

When the random variables are non-normal, the Rackwitz-Fiessler equivalent normal transformation can be used to compute the equivalent normal mean μ^e and the equivalent normal standard deviation σ^e. The iterative computations of μ^e and σ^e for each trial design point are automatic during the constrained optimization search. For non-normal random variables, Eq. (37) may be rewritten as

$$\beta_{HL} = \min_{g(\mathbf{X})=0} \sqrt{\left(\frac{\mathbf{X}-\boldsymbol{\mu}^e}{\boldsymbol{\sigma}^e}\right)^T \mathbf{R}^{-1} \left(\frac{\mathbf{X}-\boldsymbol{\mu}^e}{\boldsymbol{\sigma}^e}\right)} \tag{38}$$

where $\boldsymbol{\mu}^e$ and $\boldsymbol{\sigma}^e$ are vectors containning the equivalent mean and standard deviations values, respectively, and $\mathbf{R}^{-1}$ is the inverse of the correlation matrix. This equation will be used instead of Eq. (37) since the correlation matrix **R** displays the correlation structure more explicitly than the covariance matrix **C**.

Additional information on Solver's options and algorithms can be found in the Microsoft Excel Solver's help file and at www.solver.com. The implementation procedure of the ellipsoid approach in the spreadsheet is described in [Low97] and [Low04] among others. Some Excel files are available at http://alum.mit.edu/www/bklow.

3.4. Response Surface Method (RSM)

In case of analytically-unknown system response (such as the responses computed using a finite element/finite difference method), several approaches based on the Response Surface Method (RSM) can be found in the literature with the aim of calculating the reliability index and the corresponding design point. We present herein the approach by Tandjiria et al. [Tan00]. The basic idea of this method is to approximate the system response $Y(x)$ by an explicit function of random variables, and to improve the approximation *via* iterations. The system response can be approximated (in the original space of random variables) by:

$$Y(x) = a_0 + \sum_{i=1}^{n} a_i x_i + \sum_{i=1}^{n} b_i x_i^2 \tag{39}$$

where x_i are the random variables (μ_i and σ_i being their mean and standard deviation values, respectively); n is the number of random variables; and (a_i, b_i) are coefficients to be determined. The RSM algorithm is summarized as follows:

Step 1: Evaluate the system response $Y(x)$ at the mean value point μ and at the $2n$ points each at $\mu \pm k\sigma$ where k can be arbitrarily chosen ($k = 1$).

Step 2: The above $2n+1$ values of $Y(x)$ are used to solve eq. (39) and find the coefficients (a_i, b_i). Then, the performance function g(x) can be constructed to give a tentative response surface function.

Step 3: Solve eq. (37) to obtain a tentative design point and a tentative β_{HL} subjected to the constraint that the tentative performance function of step 2 be equal to zero.

Step 4: Repeat Steps 1 to 3 until convergence of β_{HL}. Each time, Step 1 is repeated, the $2n + 1$ sampled points are centered at the new tentative design point of Step 3.

Concerning the numerical implementation of the RSM algorithm described above, the determination of β_{HL} requires (i) the resolution of eq. (39) for the $2n+1$ sampled points, and (ii) the minimization of β_{HL} given by Eq. (37). These two operations constitute a single iteration and can be done using the optimization toolbox available in Excel. Several iterations should be performed until convergence of β_{HL}. Convergence is considered to be reached when the absolute difference between two successive values of the reliability index becomes smaller than 10^{-2}.

4. Second Order Reliability Method (SORM)

Reliability assessment is relatively simple if the limit state function is linear. However, most of the limit state functions are nonlinear. The nonlinearity is due to nonlinear relationship between random variables, to the consideration of non-normal random variables, and/or to the transformation from correlated to uncorrelated random variables. Indeed, a linear limit state in the original space becomes non-linear when transformed to the standard normal space if any of the variables is non-normal. Also, the transformation from correlated to uncorrelated variables might induce nonlinearity.

Figure 4 presents examples of linear and nonlinear limit state functions in the reduced space. Both limit state functions have the same minimum distance point β, but the failure regions are different in both cases. The failure probability of the nonlinear limit state should be less than that of the linear limit state. The FORM approach approximates the limit state function with a linear function and will therefore provide the same assessment of the probability of failure for both cases. This approximation introduces errors in the assessment of the probability of failure. Consequently, it is preferable to use a higher order approximation for the failure probability computation.

The SORM method improves the assessment given by FORM by including information about the curvature (which is related to the second-order derivatives of the limit state function with respect to the basic variables). The Taylor series expansion of a general nonlinear function $g(X_1, X_2,\ldots, X_n)$ at the design point $(x^*_1, x^*_2,\ldots, x^*_n)$ is given by:

$$g(X_1,X_2,...,X_n) = g(x_1^*,x_2^*,...,x_n^*) + \sum_{i=1}^{n} \frac{\partial g}{\partial X_i}(x_i - x_i^*) + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\partial^2 g}{\partial X_i \partial X_j}(x_i - x_i^*)(x_j - x_j^*) + ... \quad (40)$$

where the derivatives are evaluated at the design point.

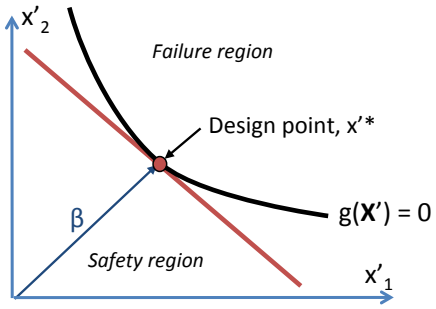


Figure 4: Linear and nonlinear limit state functions

For reliability analysis, the space of standard normal variables is more convenient for a second order approximation of $g()$. In the following, X_i and Y_i will refer to random variables in the original and equivalent uncorrelated standard normal spaces, respectively. If all the variables are uncorrelated, $Y_i = (X_i - \mu_{X_i}^e)/\sigma_{X_i}^e$ where $\mu_{X_i}^e$ and $\sigma_{X_i}^e$ are the equivalent normal mean and standard deviation of X_i at the design point x_i^*.

In the Taylor series approximation given by eq. (40), FORM ignores the terms beyond the first order term, and SORM ignores the terms beyond the second-order term (involving second-order derivatives).

Breitung [Bre84] proposed a simple closed-form solution for the probability computation using the theory of asymptotic approximation as:

$$P_f \approx \Phi(-\beta_{\text{FORM}}) \prod_{i=1}^{n-1} (1 + \beta_{\text{FORM}} \kappa_i)^{-1/2} \quad (41)$$

where κ_i represents the principal curvatures of the limit state function at the minimum distance point, and β_{FORM} is the reliability index computed by the FORM method.

The assessment of P_f requires the computation of κ_i. Towards this aim, the random variables Y_i (in the $\mathbf{Y}$ reduced space) are rotated to another set of variables Y_i', such that the last Y_i' variable coincides with the vector $\boldsymbol{\alpha}$ where $\boldsymbol{\alpha}$ is the unit gradient vector of the limit state at the minimum distance point.

Figure 5 describes the problem for two random variables indicating that the problem consists of a simply rotation of coordinates. This rotation can be carried out by an orthogonal transformation:

$$\mathbf{Y}' = \mathbf{R}\mathbf{Y} \tag{42}$$

where $\mathbf{R}$ is the rotation matrix. For instance, in the case of two random variables:

$$\mathbf{R} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \tag{43}$$

where θ is the counterclockwise angle of rotation of the axes (Figure 5). For n random variables, the reader may refer to [Hal00] for the determination of the matrix **R.**

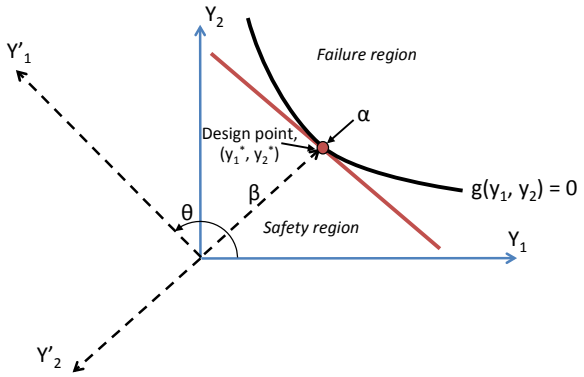


Figure 5: Rotation of axis in the standardized space

The matrix $\mathbf{R}$ is after used to estimate a matrix $\mathbf{A}$, whose elements are denoted a_{ij}, as follows:

$$a_{ij} = \frac{\left(\mathbf{R}\mathbf{D}\mathbf{R}^T\right)_{ij}}{\left|\nabla G(\mathbf{y}^*)\right|}, \qquad i,j = 1,2,...,n-1 \tag{44}$$

where $\mathbf{D}$ is the $n \times n$ second-derivative matrix of the limit state function in the standard normal space evaluated at the design point and $|\nabla G(\mathbf{y}^*)|$ is the length of the gradient vector in the standard normal space.

The last variable Y_n coincides with the β vector computed with the FORM approach in the rotated normal space. In the next step, the last row and last column in the **A** matrix and the last row in the **Y'** vector are dropped to take this factor into account. The limit state can be rewritten in terms of a second-order approximation in this rotated standard normal space **Y'** as:

$$y'_n = \beta + \frac{1}{2}\mathbf{y}'^T \mathbf{A}\mathbf{y}' \tag{45}$$

where the matrix **A** has now the size $(n-1) \times (n-1)$. Afterwards, the curvatures κ_i of eq. (41) are computed as the eigenvalues of the matrix **A**, to estimate the probability of failure.

Breitung's SORM method uses a parabolic approximation (it does not consider a general second order approximation) by ignoring the mixed terms and their derivatives in the Taylor series approximation in eq. (40). This approach uses the theory of asymptotic approximation to derive the probability estimate. This approximation is accurate for large values of β (which is the case for engineering purposes). However, the assessment is less accurate for smaller β.

Example 3 *Assessment of the reliability index by using SORM (Adapted from [San10])*

Suppose that the performance function of a problem is defined by

$$g(X_1, X_2) = X_1 X_2 - 80$$

where X_1 follows a normal distribution with mean $\mu_{X1} = 20$ and standard deviation $\sigma_{X1} = 2$. X_2 follows a lognormal distribution with mean $\mu_{X2} = 7$ and standard deviation $\sigma_{X2} = 1.4$.

*The results of the FORM approach provide a reliability index $\beta_{FORM} = 2.402$ and the two following vectors **x*** and **α**:*

$$\mathbf{x}^* = (17.612, 4.542)$$
$$\boldsymbol{\alpha} = (0.497, 0.868)$$

where ***x**** is the final design point in the original space and **α** is the vector of direction cosines.

Question 1: Estimate the probability of failure by using SORM and compare the results with the probability of failure estimated from FORM.

We already have the results of the FORM approach. We will transform X_1 and X_2 from the original to the standard normal space. Since X_2 is lognormal, we use the Rackwitz and Fiessler procedure to estimate the equivalent parameters:

$$\sigma^e_{X_2} = \frac{\phi\left[\Phi^{-1}\left[F_{X_2}\left(x_2^*\right)\right]\right]}{f_{X_2}\left(x_2^*\right)} = \frac{\phi\left[\Phi^{-1}\left[F_{X_2}\left(4.542\right)\right]\right]}{f_{X_2}\left(4.542\right)} = 0.9$$

$$\mu^e_{X_2} = x_2^* - \Phi^{-1}\left[F_{X_2}\left(x_2^*\right)\right]\sigma^e_{X_2} = 4.542 - \Phi^{-1}\left[F_{X_2}\left(4.542\right)\right]0.9 = 6.418$$

Transforming X_1 and X_2 from the original to the reduced space gives:

$$Y_1 = (X_1 - \mu_{X_1}) / \sigma_{X_1}$$

$$Y_2 = (X_2 - \mu^e_{X_2}) / \sigma^e_{X_2}$$

Then the values of the design point for each random variable in the reduced space are:

$$y_1^* = (17.612 - 20) / 2 = -1.194$$

$$y_2^* = (4.542 - 6.418) / 0.9 = -2.085$$

The first step in SORM is to determine the matrix **R** from eq. (43). In the rotated coordinates, the second coordinate need to coincide with the unit gradient vector $\boldsymbol{\alpha}$. Figure 6 presents the example in the reduced space including the performance function and the design point. It also includes the geometrical description of the assessment of θ. Thus, it can be noted that $\theta = 270^o + \tan^{-1}(-2.085/-1.194) = 330.194^o$, and **R** becomes

$$\mathbf{R} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} = \begin{bmatrix} 0.868 & -0.497 \\ 0.497 & 0.868 \end{bmatrix}$$

Notice that the elements of **R** are also easily available from the direction cosines, i.e. the components of the unit gradient vector $\boldsymbol{\alpha}$.

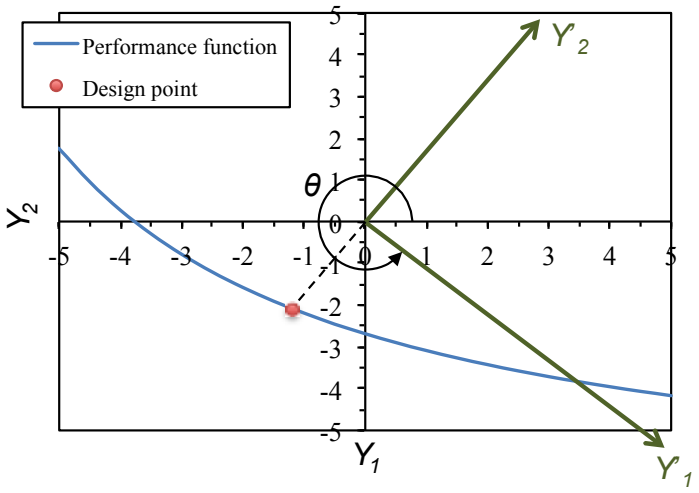


Figure 6: Rotation of the axis in the reduced space

The next step is to construct the matrix **D** that contains the second derivatives of the performance function with respect to each variable in the standard normal space. We use the chain rule of differentiation for determining the derivatives from the limit state function in the original space

$$\frac{\partial^2 g}{\partial Y_1^2} = \frac{\partial}{\partial X_1}\left[\frac{\partial g}{\partial X_1}\frac{\partial X_1}{\partial Y_1}\right]\frac{\partial X_1}{\partial Y_1} = \frac{\partial}{\partial X_1}\left[X_2\sigma_{X_1}\right]\sigma_{X_1} = 0$$

$$\frac{\partial^2 g}{\partial Y_2^2} = \frac{\partial}{\partial X_2}\left[\frac{\partial g}{\partial X_2}\frac{\partial X_2}{\partial Y_2}\right]\frac{\partial X_2}{\partial Y_2} = \frac{\partial}{\partial X_2}\left[X_1\sigma^e_{X_2}\right]\sigma_{X_2} = 0$$

$$\frac{\partial^2 g}{\partial Y_1 Y_2} = \frac{\partial}{\partial X_1}\left[\frac{\partial g}{\partial X_2}\frac{\partial X_2}{\partial Y_2}\right]\frac{\partial X_1}{\partial Y_1} = \frac{\partial}{\partial X_1}\left[X_1\sigma^e_{X_2}\right]\sigma_{X_1} = \sigma^e_{X_2}\sigma_{X_1}$$

Consequently

$$\mathbf{D} = \begin{bmatrix} \frac{\partial^2 g}{\partial Y_1^2} & \frac{\partial^2 g}{\partial Y_1 Y_2} \\ \frac{\partial^2 g}{\partial Y_1 Y_2} & \frac{\partial^2 g}{\partial Y_2^2} \end{bmatrix} = \begin{bmatrix} 0 & \sigma^e_{X_2}\sigma_{X_1} \\ \sigma^e_{X_2}\sigma_{X_1} & 0 \end{bmatrix} = \begin{bmatrix} 0 & 1.799 \\ 1.799 & 0 \end{bmatrix}$$

We estimate the following partial derivatives for the assessment of **A**

$$\frac{\partial g}{\partial Y_1} = \frac{\partial g}{\partial X_1}\frac{\partial X_1}{\partial Y_1} = X_2\sigma_{X_1}$$
$$\frac{\partial g}{\partial Y_2} = \frac{\partial g}{\partial X_2}\frac{\partial X_2}{\partial Y_2} = X_1\sigma^e_{X_2}$$

These derivatives evaluated at the design point are

$$\nabla G(\mathbf{y}^*) = \begin{bmatrix} x_2^*\sigma_{X_1} \\ x_1^*\sigma^e_{X_2} \end{bmatrix} = \begin{bmatrix} (17.612)(2) \\ (4.542)(0.9) \end{bmatrix} = \begin{bmatrix} 35.223 \\ 4.086 \end{bmatrix}$$

and the length of the vector is

$$\left|\nabla G(\mathbf{y}^*)\right| = \sqrt{35.223^2 + 4.086^2} = 35.46$$

Therefore by applying eq. (44), the matrix **A** is

$$\mathbf{A} = \begin{bmatrix} -0.044 & 0.026 \\ 0.026 & 0.044 \end{bmatrix}$$

Afterwards, we modify the matrix **A** by deleting the last row and column. In this case, we have only one element $a_{11} = -0.044$. Then, the eigenvalue of this one element matrix is simply the same element: $\kappa_1 = a_{11} = -0.044$.

The approximation of the failure probability is estimated from eq. (41)

$$P_f \approx \Phi(-\beta_{\text{FORM}})\prod_{i=1}^{n-1}(1+\beta_{\text{FORM}}\kappa_i)^{-1/2}$$
$$\approx \Phi(-2.402)\frac{1}{\sqrt{1+(2.402)(-0.044)}} \approx 8.609\cdot 10^{-3}$$

The probability of failure estimated from FORM is $P_f = \Phi(-2.402) = 8.144\cdot 10^{-3}$. The probability of failure estimate from the SORM procedure is about 5% larger. This result is expected because of the curvature of the performance function (Figure 6).

5. Conclusions

This chapter presented the First Order Reliability Method (FORM) and the Second Order Reliability Method (SORM). The First Order Reliability Method (FORM) makes use of the first and second moments of the random variables. This method includes two approaches. These are First-Order Second-Moment (FOSM) and Ad-

vanced First-Order Second-Moment (AFOSM) approaches. In FOSM, the information on the distribution of random variables is ignored; however, in AFOSM (called also Hasofer-Lind approach), the distributional information is appropriately used. It was shown that contrary to FOSM, the Hasofer-Lind method led to an invariant reliability index regardless of the form in which the limit state equation is written.

While AFOSM requires the transformation of the limit state surface to a standard space of random variables, the recent ellipsoid approach led to a simple method for computing the Hasofer-Lind reliability index in the original space of random variables using an optimization tool available in most spreadsheet software packages.

In case of analytically-unknown system response (i.e. when the system response is computed using a finite element/finite difference method), the Response Surface Method (RSM) can be used to calculate the Hasofer-Lind reliability index and the corresponding design point. The basic idea of this method is to approximate the system response by an explicit function of random variables, and to improve the approximation via iterations.

For the computation of the failure probability, it was shown that the Hasofer-Lind reliability index can be used to evaluate the failure probability when the limit state function is a linear function of uncorrelated normal variables or when the nonlinear limit state function is represented by a first-order (linear) approximation with equivalent normal variables. SORM estimates the probability of failure by approximating the nonlinear limit state function (including a linear limit state function with correlated non-normal variables) by a second-order representation.

References

[Bre84] Breitung, K. Asymptotic approximations for multinormal integrals. *J. Eng. Mech., ASCE*, *110*(3), 357–367, 1984.

[Cor69] Cornell, C. A. Bounds on the reliability of structural systems. *J. struct. div., ASCE*, *93*(1), 171-200, 1967.

[Low97] Low, B. K., and Tang, W. H. Efficient reliability evaluation using spreadsheet. *J. Eng. Mech.,* 123(7), 749–752, 1997.

[Low04] Low, B. K., and Tang, W. H. Reliability analysis using object-oriented constrained optimization. *Struct. Safety,* 26, 68–89.

[Hal00] Haldar A. and Mahadevan S. *Probability, Reliability and Statistical Methods in Engineering desing*, Jonh Wiley & Sons, Inc. 2000..

[Rac76] Rackwitz, R. *Practical probabilistic approach to design,* Bulletin No. 112, CEB, Paris France, 1976.

[Rac78] Rackwitz, R. and Fiessler R. Structural reliability under combined random load sequences, *Structural Safety,* 22(1), 27–60, 1978.

[San10] Sanchez-Silva, M. *Introduction to reliability and risk assessment: theory and applications in engineering,* Ediciones Uniandes, 2010. (In Spanish).

[Tan00] Tandjiria, V., Teh, C.I. and Low, B.K. Reliability analysis of laterally loaded piles using response surface methods, Struct Saf 22:335–355, 2000.

Appendix: Summary of Matlab® functions for the normal distribution

Function	Description
normpdf(X,mu,sigma)	computes the PDF at each of the values in X using the normal distribution with mean mu and standard deviation sigma. X, mu, and sigma can be vectors, matrices, or multidimensional arrays that all have the same size. A scalar input is expanded to a constant array with the same dimensions as the other inputs. The parameters in sigma must be positive.
normcdf(X)	returns the standard normal CDF at each value in X. The standard normal distribution has parameters mu = 0 and sigma = 1.
norminv(P,mu,sigma)	computes the inverse of the normal CDF using the corresponding mean mu and standard deviation sigma at the corresponding probabilities in P. P, mu, and sigma can be vectors, matrices, or multidimensional arrays that all have the same size. A scalar input is expanded to a constant array with the same dimensions as the other inputs. The parameters in sigma must be positive, and the values in P must lie in the interval [0 1].

Advanced reliability analysis methods

Abdul-Hamid Soubra and Emilio Bastidas-Arteaga

University of Nantes – GeM Laboratory, France

This chapter presents the subset simulation (SS) approach and the Polynomial Chaos Expansion (PCE) methodology. The SS method is an efficient alternative to the well-known Monte Carlo Simulation (MCS) methodology to calculate small failure probabilities. The basic idea of the SS approach is that the small failure probability can be expressed as a product of larger conditional failure probabilities. On the other hand, the PCE methodology allows one to accurately compute the PDF of a given system response using a reduced number of calls of the deterministic model (as compared to the classical MCS applied on the original complex deterministic model). Indeed, the PCE methodology replaces the computationally-expensive deterministic model by a meta-model. Once the meta-model is determined, MCS can be applied on the obtained PCE to compute the PDF of the system response with a quasi-negligible computation time.

1 Introduction

The most robust method used for the probabilistic analysis of geotechnical structures is the classical well-known Monte Carlo Simulation (MCS) methodology. It should be noted that the probabilistic analysis of an engineering system involves the computation of the PDF of the system response or the calculation of the failure probability for a prescribed threshold of this system response.

MCS is not suitable for the computation of the small failure probabilities encountered in the practice of geotechnical engineering (especially when using a computationally-expensive finite element/finite difference deterministic model) due to the large number of simulations required to calculate a small failure probability. As an alternative to MCS methodology, [Au01] proposed the subset simulation (SS) approach to calculate small failure probabilities. The basic idea of the SS approach is that the small failure probability can be expressed as a product of larger conditional failure probabilities.

Similarly, MCS is not suitable for the accurate determination of the PDF of a system response because of the great number of calls of the deterministic model, which are

required for such a computation. The PCE methodology allows one to approximate a given system response by a polynomial chaos expansion (PCE) of a suitable order. Thus, the PCE methodology replaces the computationally-expensive deterministic model by a meta-model (i.e. a simple analytical equation). Once the PCE coefficients are determined, MCS can be applied on the obtained PCE to compute the PDF of the system response (and the corresponding statistical moments) with a quasi-negligible computation time.

2 Subset simulation (SS) approach

Subset simulation (SS) approach was proposed by [Au01] as alternative to Monte Carlo Simulation (MCS) methodology to compute small failure probabilities. The basic idea of the SS approach is that the small failure probability can be expressed as a product of larger conditional failure probabilities. In this section, one presents a brief description of the steps of SS approach in case of two random variables, the extension to the case of several random variables being straightforward. A detailed description of the SS approach may be found in [Au01], in the chapters 1 and 4 of the book by [Pho08] and in [Ahm12].

The steps of SS approach in case of two random variables (V_1, V_2) can be described as follows:

1. Generate a vector of two random variables (V_1, V_2) according to a target PDF using direct Monte Carlo simulation.
2. Using the deterministic model, calculate the system response corresponding to (V_1, V_2).
3. Repeat steps 1 and 2 until obtaining a prescribed number N_s of vectors of random variables and the corresponding values of the system response.
4. Determine the value of the performance function corresponding to each value of the system response and then, arrange the values of the performance function in an increasing order within a vector G_0 where $G_0 = \{G_0^1, \dots, G_0^k, \dots, G_0^{N_s}\}$. Notice that the subscripts '*0*' refer to the first level (level *0*) of the subset simulation approach.
5. Prescribe a constant intermediate conditional failure probability p_0 for the failure regions $F_j\{j = 1,2, \dots, m-1\}$ and evaluate the first failure threshold C_1 which corresponds to the first level of SS approach (see Figure 1). The failure threshold C_1 is equal to the $[(N_s \times p_0) + 1]^{th}$ value in the increasing list of elements of the vector G_0. This means that the value of the conditional failure probability of the first level $P(F_1)$ will be equal to the prescribed p_0 value.
6. Among the N_s vectors of random variables, there are $[N_s \times p_0]$ ones whose values of the performance function are less than C_1 (i.e. they are located in the failure region F_1). These vectors are used as 'mother vectors' to generate N_s new vectors of random variables (according to a proposal P_p) using Markov

chain method based on the modified Metropolis-Hastings algorithm by [San11]. This algorithm is presented in Appendix 1.

7. Using the deterministic model, calculate the values of the system response corresponding to the new vectors of random variables (which are located in level 1). Then, calculate the corresponding values of the performance function. Finally, gather the values of the performance function in an increasing order within a vector G_1 where $G_1 = \{G_1^1, \dots, G_1^k, \dots, G_1^{N_s}\}$.
8. Evaluate the second failure threshold C_2 as the $[(N_s \times p_0) + 1]^{th}$ value in the increasing list of elements of the vector G_1.
9. Repeat steps 6-8 to evaluate the failure thresholds $C_3, C_4, \dots, C_m$ corresponding to the failure regions $F_3, F_4, \dots, F_m$. Notice that contrary to all other thresholds, the last failure threshold C_m is negative. Thus, C_m is set to zero and the conditional failure probability of the last level $P(F_m|F_{m-1})$ is calculated as:

$$P(F_m|F_{m-1}) = \frac{1}{N_s}\sum_{k=1}^{N_s} I_{F_m(s_k)} \tag{1}$$

where $I_{F_m} = 1$ if the performance function $G(s_k)$ is negative and $I_{F_m} = 0$ otherwise.

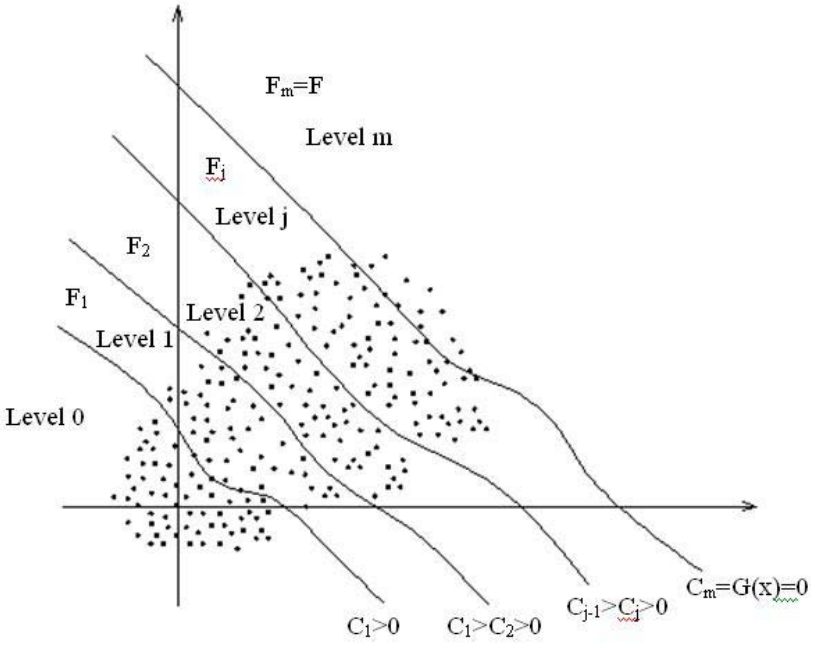


Figure 1: Nested Failure domain.

Finally, the failure probability $P(F)$ is evaluated as follows:

$$P(F) = P(F_1)\prod_{j=2}^{m} P\left(F_j \middle| F_{j-1}\right) \tag{2}$$

It should be mentioned that a normal PDF was used herein as a target probability density function P_t. However, a uniform PDF was used as a proposal probability density function P_p (for more details, refer to Appendix 1).

2.1 Example application

In order to illustrate the algorithm of SS methodology in a simple way, a numerical example is provided herein. In this example, SS approach was used to calculate the failure probability P_f against bearing capacity failure of a strip footing of breadth B. The footing rests on a (c, φ) soil and it is subjected to a service vertical load P_s. The soil cohesion c and the soil angle of internal friction φ were considered as random variables. The following formula was used for the computation of the ultimate bearing capacity:

$$q_u = \gamma\frac{B}{2}N_\gamma + cN_c + qN_q \tag{3}$$

in which:

$$N_\gamma = 2\left(N_q - 1\right)tan\varphi \tag{4a}$$

$$N_q = e^{\pi tan\varphi}.tan^2\left(\frac{\pi}{4} + \frac{\varphi}{2}\right) \tag{4b}$$

$$N_c = \frac{N_q - 1}{tan\varphi} \tag{4c}$$

where N_γ, N_q and N_c are the bearing capacity factors due to soil weight, surcharge loading and cohesion respectively. These coefficients are function of the soil friction angle. On the other hand, γ is the soil unit weight and q is the surcharge loading. The performance function used in the analysis is:

$$G = \frac{P_u}{P_s} - 1 \tag{5}$$

where P_u is the ultimate footing load and P_s is the footing applied load. As mentioned previously, only the soil cohesion and friction angle were considered as random variables. All the other parameters were considered as deterministic. These parameters are given in Table (1).

In this example, the intermediate failure probability p_0 of a given level j ($j = 1,2,\dots,m-1$) was arbitrary chosen equal to 0.2. A small number of samples per level ($N_s = 10$ samples) was used to facilitate the illustration.

Table 1: Data used for the probabilistic analysis of a strip footing against bearing capacity failure

Parameter	**Type of parameter**	**Mean and coefficient of variation of the parameter**
Breadth B	Deterministic	2m
Surcharge loading q	Deterministic	10kPa
Soil unit weight γ	Deterministic	20kN/m^3
Service vertical load P_s	Deterministic	1000kN/m
Cohesion c	Random normal variable	μ_c = 20kPa
		COV_c= 0.3
Friction angle φ	Random normal variable	μ_φ = 30°
		COV_φ = 0.1

Table 2 presents (i) the values of c and φ of each sample for the successive levels (ii) the corresponding values of the performance function and (iii) the values of the failure thresholds C_j for the different levels. Notice that only the first two levels and the last level for which the failure threshold becomes negative were provided herein for illustration. Table (2) indicates that the failure threshold decreases with the successive levels until reaching a negative value at the last level. This means that the samples generated by the subset simulation successfully progress towards the limit state surface $G = 0$. In order to select the failure threshold of a given level, the calculated values of the performance function of this level were arranged in an increasing order as shown in Table (2). Then, the failure threshold was selected as the $[(N_s \times p_0) + 1]^{th}$ value of the arranged values of the performance function. Since $N_s = 10$ and $p_0 = 0.2$, the failure threshold is equal to the third value of the arranged values of the performance function. The SS computation continues until reaching a negative value (or a value of zero) of the failure threshold. In this example, the negative value was reached in the sixth level (where $C_6 = -0.0936$) as shown in Table (2). Theoretically, the last failure threshold should be equal to zero. For this reason, C_6 was set to zero. This means that the last conditional failure probability $P(F_6|F_5)$ is not equal to p_0. In this case, the last conditional failure probability $P(F_6|F_5)$ is calculated as the ratio between the number of samples for which the performance function is negative and the chosen number N_s of samples (i.e. 10). According to Table (2), $P(F_6|F_5)$ is equal to $3/10 = 0.3$. Thus, the failure probability of the footing under consideration is equal to $0.2^5 \times 0.3 = 9.6 \times 10^{-5}$.

Table 2: Results of SS algorithm when N_s=10 and p_0=0.2

Level's number j	Cohesion c (kPa)	Angle of internal friction φ (deg)	Performance function	Failure threshold C_j
1	23.23	26.0	0.9256	1.4875
	31.00	39.1	1.1185	
	6.45	32.2	1.4875	
	25.17	29.9	1.5598	
	21.91	32.1	2.0023	
	12.15	29.4	2.6625	
	17.40	29.6	3.8598	
	22.06	34.5	4.9894	
	41.47	34.3	5.3910	
	36.62	34.3	9.1912	
2	25.83	26.5	0.8587	0.9740
	27.29	26.2	0.9411	
	25.32	26.8	0.9740	
	23.98	25.4	1.0561	
	25.92	26.0	1.0842	
	14.86	28.0	1.1505	
	14.14	28.8	1.1528	
	12.27	28.8	1.1747	
	13.14	29.4	1.1931	
	11.80	30.0	1.2361	
6	15.17	22.9	-0.1604	-0.0936
	14.88	23.0	-0.1003	
	14.88	23.0	-0.0936	
	14.56	22.5	0.0415	
	14.56	22.5	0.0718	
	15.84	22.5	0.0718	
	16.36	21.5	0.1156	
	14.53	20.7	0.1420	
	12.89	20.6	0.1476	
	15.43	20.3	0.1476	
P_f	9.6×10^{-5}			

It should be emphasized that the failure probability calculated in Table (2) is not accurate due to the small value of N_s. For an accurate computation of the failure probability, N_s should be increased. This number should be greater than 100 to provide a small bias in the calculated P_f value (see chapter 4 by Honjo in [Pho08]).

In order to determine the optimal number of samples N_s to be used per level, different values of N_s were considered to calculate P_f and its coefficient of variation COV_{P_f} as shown in Table (3). The thresholds corresponding to each N_s value were calculated and were shown in this table. Table (3) indicates (as was shown before when $N_s = 10$) that for the different values of N_s, the failure threshold decreases with the successive levels until reaching a negative value at the last level.

Table 3: Evolution of the failure threshold C_j with the different levels j and with the number of realizations N_s when $p_0 = 0.2$

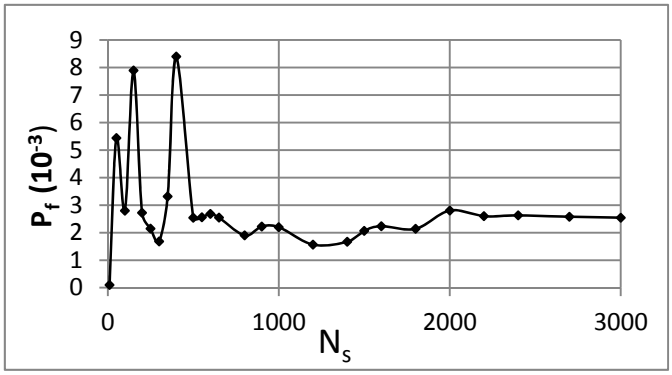
C_j for level j	Number of samples N_s per level						
	10	100	200	1000	2000	2200	2400
C_1	1.4875	0.9397	1.0071	1.0638	1.0532	1.0466	1.0803
C_2	0.9740	0.4157	0.3969	0.4916	0.4467	0.4466	0.4942
C_3	0.7391	0.1011	0.1016	0.1513	0.1434	0.1347	0.1549
C_4	0.4007	-0.0491	-0.0437	-0.0307	-0.0616	-0.0536	-0.0564
C_5	0.1573	--------	--------	--------	--------	--------	--------
C_6	-0.0936	--------	--------	--------	--------	--------	--------
$P_f(\times 10^{-3})$	0.096	2.80	2.72	2.20	2.80	2.60	2.63
$COV_{P_f}(\%)$	221.4	57.9	42.1	18.7	13.3	12.8	12.4

Figure (2a) shows the effect of N_s on the failure probability. It indicates that for small values of N_s, the failure probability largely changes with N_s. However, for high values of N_s, the failure probability converges to an almost constant value. Figure (2a) also indicates that 2200 samples per level are required to accurately calculate the failure probability. This is because (i) the C_j values corresponding to N_s=2200 and 2400 samples are quasi similar as it may be seen from Table (3) and (ii) the corresponding final P_f values are too close (they are respectively equal to 2.60×10^{-3} and 2.63×10^{-3}).

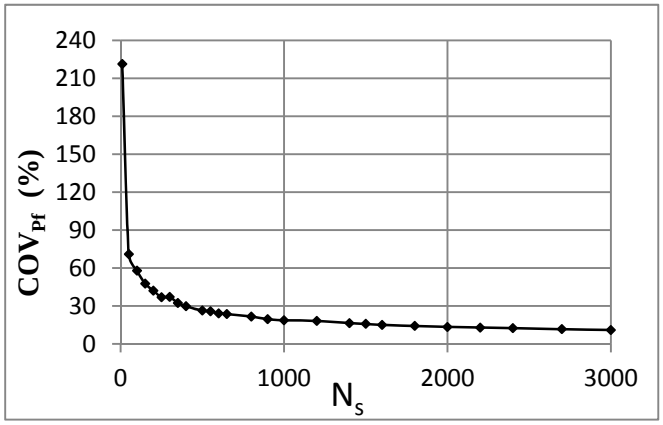
Figure (2b) shows the effect of N_s on the coefficient of variation of the failure probability COV_{P_f}. As expected, COV_{P_f} decreases with the increase of N_s. Notice that the values of COV_{P_f} for N_s =2200 and 2400 samples are equal to 12.8% and 12.4% which indicates (as expected) that the COV_{P_f} decreases with the increase in the number of realizations.

It should be mentioned here that for $p_0 = 0.2$, four levels of subset simulation were found necessary to reach the limit state surface $G = 0$ as may be seen from Table (3). Therefore, when N_s=2200 samples, a total number of N_t=*2200×4=8800* samples were required to calculate the final P_f value. Remember that in this case, the COV of P_f was equal to *12.8%*. Notice that if the same value of COV (*i.e. 12.8%*) is

desired by MCS to calculate P_f, the number of samples would be equal to *20000*. This means that, for the same accuracy, the SS approach reduces the number of realizations by 56%. On the other hand, if one uses MCS with the same number of samples (i.e. *8800* realizations), the value of COV of P_f would be equal to *19.6%*. This means that for the same computational effort, the SS approach provides a smaller value of $COV(P_f)$ than MCS.



(a)



(b)

Figure 2: P_f and COV_{P_f} versus the number of realizations N_s.

3. Polynomial Chaos Expansion (PCE) methodology

The basic idea of this method is to approximate a given system response by a polynomial chaos expansion (PCE) of a suitable order. In other words, the PCE methodology replaces the computationally-expensive deterministic model by a meta-model.

In order to achieve this purpose, all the uncertain parameters (which may have different PDFs) should be represented by a unique chosen PDF. Table (4) presents the usual PDFs and their corresponding families of orthogonal polynomials.

Table 4: Usual probability density functions and their corresponding families of orthogonal polynomials

Probability density functions	**Polynomials**
Gaussian	Hermite
Gamma	Laguerre
Beta	Jacobi
Uniform	Legendre

Within the framework of the present methodology, the response of a system that involves *n* random variables can be expressed by a PCE as follows:

$$\Gamma_{PCE} = \sum_{i=0}^{P-1} a_i \Psi_i(\xi) \tag{6}$$

where $\psi_i(\xi)$ are multi-dimensional polynomials defined as the product of one-dimensional polynomials ϕ_{α_i}, $(\xi_1, \xi_2, \ldots, \xi_n)$ are independent random variables, $(a_1, a_2, \ldots, a_n)$ are unknown coefficients to be evaluated and *P* is the size of the PCE.

The size *P* of the PCE (which is equal to the number of the unknown PCE coefficients) depends on the number *n* of random variables and the order *p* of the PCE. It is given as follows:

$$P = \frac{(n+p)!}{n!\,p!} \tag{7}$$

It should be mentioned here that in this chapter, the random variables are represented in the independent standard normal space. Thus, the suitable corresponding bases are the multidimensional Hermite polynomials as may be seen from Table (4). The expressions of the multi-dimensional Hermite polynomials are given as follows:

$$\psi_\alpha = \prod_{i=1}^{n} \phi_{\alpha_i}(\xi_i) \tag{8}$$

where $\alpha=[\alpha_1, \ldots, \alpha_n]$ is a sequence of *n* non-negative integers and $\phi_{\alpha_i}(\xi_i)$ are one-dimensional Hermite polynomials. More details on the one-dimensional Hermite polynomials are given in Appendix 2.

For the determination of the PCE unknown coefficients, a non-intrusive technique (in which the deterministic model is treated as a black-box) is used (see [Ahm12] among others). In this chapter, the regression approach is employed. In this approach, it is required to compute the system response at a set of collocation points in order to perform a fit of the PCE using the obtained system response values.

As suggested by several authors (e.g. [Hua09]), the collocation points can be chosen as the result of all possible combinations of the roots of the one-dimensional Hermite polynomial of order $(p+1)$ for each random variable. For example, if a PCE of order p=2 is used to approximate the response surface of a system with n=2 random variables, the roots of the one-dimensional Hermite Polynomial of order 3 are chosen for each random variable. These roots are (-√3, 0, √3) for the first random variable and (-√3, 0, √3) for the second random variable. In this case, 9 collocation points are available. These collocation points are (-√3, -√3), (-√3, 0), (-√3, √3), (0, -√3), (0, 0), (0, √3), (√3, -√3), (√3, 0), (√3, √3). In the general case, for a PCE of order p and for n random variables, the number N of the available collocation points can be obtained using the following formula:

$$N=(p+1)^n \tag{9}$$

Referring to Equations (7 and 9), one can observe that the number of the available collocation points is higher than the number of the unknown coefficients. This leads to a linear system of equations whose number N of equations is greater than the number P of the unknown coefficients. The regression approach is used to solve this system. This approach is based on a least square minimization between the exact solution Γ and the approximate solution Γ_{PCE} which is based on the PCE. Accordingly, the unknown coefficients of the PCE can be computed using the following equation:

$$\boldsymbol{a} = (\boldsymbol{\Psi}^T \boldsymbol{\Psi})^{-1}.\ \boldsymbol{\Psi}^T.\ \boldsymbol{\Gamma} \tag{10}$$

in which $\boldsymbol{a}$ is a vector containing the PCE coefficients, $\boldsymbol{\Gamma}$ is a vector containing the system response values as calculated by the deterministic model at the different collocation points and $\boldsymbol{\Psi}$ is a matrix of size $N \times P$ whose elements are the multivariate Hermite polynomials. It is given as follows:

$$\boldsymbol{\Psi} = \begin{bmatrix} \psi_0^1(\xi) & \psi_1^1(\xi) & \psi_2^1(\xi) & \dots & \psi_{P-1}^1(\xi) \\ \psi_0^2(\xi) & \psi_1^2(\xi) & \psi_2^2(\xi) & \dots & \psi_{P-1}^2(\xi) \\ \vdots & \vdots & \vdots & & \vdots \\ \vdots & \vdots & \vdots & & \vdots \\ \psi_0^N(\xi) & \psi_1^N(\xi) & \psi_2^N(\xi) & \dots & \psi_{P-1}^N(\xi) \end{bmatrix} \tag{11}$$

Notice that in order to calculate the system response corresponding to a given collocation point, the standard normal random variables ξ_i should be expressed in the original physical space of random variables as follows:

$$x_i = F_{x_i}^{-1}\left[\Phi(\xi_i)\right] \tag{12}$$

in which, x_i is a physical random variable, F_{xi} is the CDF of the physical random variable and Φ is the CDF of the standard normal random variable. Notice also that if the original physical random variables are correlated, the standard normal random variables should first be correlated using the following equation:

$$\begin{bmatrix} \xi_{1c} \\ \xi_{2c} \\ . \\ . \\ . \\ \xi_{nc} \end{bmatrix} = H. \begin{bmatrix} \xi_1 \\ \xi_2 \\ . \\ . \\ . \\ \xi_n \end{bmatrix} \tag{13}$$

in which $\{\xi_{1c}, \xi_{2c}, ..., \xi_{nc}\}$ is the vector of correlated standard normal random variables, $\{\xi_1, \xi_2, ..., \xi_n\}$ is the vector of uncorrelated standard normal random variables and H is the Cholesky transformation of the correlation matrix of the physical random variables.

Once the PCE coefficients are determined, MCS can be applied on the obtained PCE (called meta-model) to compute the PDF of the system response. This is achieved by (i) generating a large number of realizations of the vector $(\xi_1, \xi_2, ... \xi_n)$ of standard normal random variables and (ii) calculating the system response corresponding to each realization by substituting the vector $(\xi_1, \xi_2, ... \xi_n)$ in the meta-model.

3.1. Optimal number of collocation points

The number of the available collocation points significantly increases with the increase in the number of random variables (cf. Eq. 9) and it may be very large with respect to the number of the unknown PCE coefficients. This makes it necessary to determine the optimal number of collocation points which is needed by the regression approach to solve the linear system of equations (Eq. 10). Sudret [Sud08] has proposed to successively increase the information matrix $\boldsymbol{A}$ [where $\boldsymbol{A}=(\boldsymbol{\Psi}^T\boldsymbol{\Psi})$] until it becomes invertible as follows: (a) the N available collocation points are ordered in a list according to increasing norm, (b) the information matrix $\boldsymbol{A}$ is constructed using the first P collocation points of the ordered list, i.e. the P collocation points that are the closest ones to the origin of the standard space of random variables and finally

(c) this matrix is successively increased (by adding each time the next collocation point from the ordered list) until it becomes invertible.

3.2. Accuracy of the obtained PCE

For a given PCE order, the accuracy of the approximation of the system response by a PCE can be measured by the coefficient of determination. Two types of coefficients of determination exist in literature. These are the coefficient of determination R^2 and the *leave-one-out* coefficient of determination Q^2.

Let us consider J realizations of the standard normal random vector ξ as follows: $\left\{ \xi^{(1)} = \left(\xi_1^{(1)}, \ldots, \xi_n^{(1)}\right), \ldots, \xi^{(J)} = \left(\xi_1^{(J)}, \ldots, \xi_n^{(J)}\right) \right\}$, and let us assume that the vector $\Gamma = \left\{ \Gamma\left(\xi^{(1)}\right), \ldots, \Gamma\left(\xi^{(J)}\right) \right\}$ includes the corresponding values of the system response determined by deterministic calculations. The coefficient of determination R^2 is calculated as follows:

$$R^2 = 1 - \Delta_{PCE} \tag{14}$$

where Δ_{PCE} is given by:

$$\Delta_{PCE} = \frac{(1/J)\sum_{i=1}^{J}\left[\Gamma\left(\xi^{(i)}\right) - \Gamma_{PCE}\left(\xi^{(i)}\right)\right]^2}{Var(\Gamma)} \tag{15}$$

and

$$Var(\Gamma) = \frac{1}{J-1}\sum_{i=1}^{J}\left[\Gamma\left(\xi^{(i)}\right) - \bar{\Gamma}\right]^2 \tag{16}$$

Note here that J is the number of collocation points used to evaluate the unknown coefficients of the PCE. The value $R^2 = 1$ indicates a perfect approximation of the true system response Γ, whereas $R^2 = 0$ indicates a nonlinear relationship between the true model Γ and the PCE model Γ_{PCE}.

The coefficient of determination R^2 may be a biased estimate since it does not take into account the robustness of the meta-model (i.e. its capability of correctly predicting the model response at any point which does not belong to the collocation points. As a consequence, a more reliable and rigorous coefficient of determination, called the *leave-one-out* coefficient of determination, was proposed in literature. This coefficient of determination consists in sequentially removing a point from the J collocation points. Let $\Gamma_{\tilde{\xi}/i}$ be the meta-model that has been built from $(J-1)$ collocation points after removing the i^{th} observation from these collocation points and let $\Delta^i = \Gamma(\xi^{(i)}) - \Gamma_{\tilde{\xi}/i}(\xi^{(i)})$ be the predicted residual between the model evaluation at

point $\xi^{(i)}$ and its prediction at the same point based on $\Gamma_{\xi/i}$. The empirical error is thus given as follows:

$$\Delta^*_{PCE} = \frac{1}{J}\sum_{i=1}^{J}(\Delta^i)^2 \tag{17}$$

The corresponding coefficient of determination is often denoted by Q^2 and is called *leave-one-out* coefficient of determination. It is given as follows:

$$Q^2 = 1 - \frac{\Delta^*_{PCE}}{Var(\Gamma)} \tag{18}$$

3.3. PCE-based Sobol indices

A Sobol index of a given input random variable is a measure by which the contribution of this input random variable to the variability of the system response can be determined. Sobol indices are generally calculated by MCS methodology. This method is very time-expensive especially when dealing with a large number of random variables. [Sud08] proposed an efficient approach to calculate the Sobol indices based on the coefficients of the PCE. This method is based on ranking the different terms of the PCE and gathering them into groups where each group contains only one random variable or a combination of random variables. For more details on the computation of Sobol indices, the reader may refer to [Ahm12] among others.

4. Conclusion

This chapter first presented the subset simulation approach which is an efficient alternative to MCS for the computation of a small failure probability. An example application was provided. It aims at showing the practical implementation of the SS approach. It was found that for a prescribed accuracy, the SS approach significantly reduces the number of realizations as compared to Monte Carlo simulations methodology (the reduction was found equal to 93.3% in the present chapter). In other words, for the same computational effort, the SS approach provides a smaller value of the coefficient of variation of P_f than MCS. It should be mentioned that the Matlab code used for the example application is provided in http://www.univ-nantes.fr/soubra-ah for practical use.

In a second stage, the Polynomial Chaos Expansion methodology was presented. It was shown that the PCE method replaces the computationally-expensive deterministic model by a meta-model (i.e. a simple analytical equation). Once the PCE coefficients are determined, MCS can be applied on the obtained PCE to easily compute the PDF of the system response with a quasi-negligible computation time.

APPENDIX 1

Modified METROPOLIS-HASTINGS algorithm

The Metropolis–Hastings algorithm is a Markov chain Monte Carlo ($MCMC$) method. It is used to generate a sequence of new realizations from existing realizations (that follow a target PDF called 'P_t'). Refer to Figure (1) and let $s_k \in F_j$ be a current realization which follows a target PDF 'P_t'. Using a proposal PDF 'P_p', a next realization $s_{k+1} \in F_j$ that follows the target PDF 'P_t' can be simulated from the current realization s_k as follows:

a. A candidate realization $\hat{s}$ is generated using the proposal PDF (P_p). The candidate realization $\hat{s}$ is centered at the current realization s_k.
b. Using the deterministic model, evaluate the value of the performance function $G(\hat{s})$ corresponding to the candidate realization $\hat{s}$. If $G(\hat{s}) < C_j$ (i.e. $\hat{s}$ is located in the failure region F_j), set $s_{k+1} = \hat{s}$; otherwise, reject $\hat{s}$ and set $s_{k+1} = s_k$ (i.e. the current realization s_k is repeated).
c. If $G(\hat{s}) < C_j$ in the preceding step, calculate the ratio r_1=$P_t(\hat{s})/P_t(s_k)$ and the ratio r_2=$P_p(s_k|\hat{s})/P_p(\hat{s}|s_k)$, then compute the value $r = r_1 r_2$.
d. If $r \geq 1$ (i.e. $\hat{s}$ is distributed according to the P_t), one continues to retain the realization s_{k+1} obtained in step b; otherwise, reject $\hat{s}$ and set $s_{k+1} = s_k$ (i.e. the current realization s_k is repeated).

Notice that in step b, if the candidate realization $\hat{s}$ does not satisfy the condition $G(\hat{s}) < C_j$, it is rejected and the current realization s_k is repeated. Also in step d, if the candidate realization $\hat{s}$ does not satisfy the condition $r \geq 1$ (i.e. $\hat{s}$ is not distributed according to the P_t), it is rejected and the current realization s_k is repeated. The presence of several repeated realizations is not desired as it leads to high probability that the chain of realizations remains in the current state. This means that there is high probability that the next failure threshold C_{j+1} is equal to the current failure threshold C_j. This decreases the efficiency of the subset simulation approach. To overcome this inconvenience, Santoso *et al*. (2011) proposed to modify the classical M-H algorithm as follows:

a. A candidate realization $\hat{s}$ is generated using the proposal (P_p). The candidate realization $\hat{s}$ is centered at the current realization s_k.
b. Calculate the ratio r_1=$P_t(\hat{s})/P_t(s_k)$ and the ratio r_2=$P_p(s_k|\hat{s})/P_p(\hat{s}|s_k)$, then compute the value $r = r_1 r_2$.
c. If $r \geq 1$, set $s_{k+1} = \hat{s}$; otherwise, another candidate realization is generated. Candidate realizations are generated randomly until the condition $r \geq 1$ is satisfied.
d. Using the deterministic model, evaluate the value of the performance function $G(s_{k+1})$ of the candidate realization that satisfies the condition $r \geq 1$. If $G(s_{k+1}) < C_j$ (i.e. s_{k+1} is located in the failure region F_j), one continues

to retain the realization s_{k+1} obtained in step c; otherwise, reject $\hat{s}$ and set $s_{k+1} = s_k$ (i.e. the current realization s_k is repeated).

These modifications reduce the repeated realizations and allow one to avoid the computation of the system response of the rejected realizations. This becomes of great importance when the time cost for the computation of the system response is expensive (i.e. for the finite element or finite difference models).

APPENDIX 2

The one-dimensional Hermite polynomials of orders 0, 1, 2, 3, …, p+1 are given by:

$\phi_0(\xi) = 1$

$\phi_1(\xi) = \xi$

$\phi_2(\xi) = \xi^2 - 1$

$\phi_3(\xi) = \xi^3 - 3\xi$

$\phi_{p+1}(\xi) = \xi\,\phi_p(\xi) - p\phi_{p-1}(\xi)$

where ξ is a standard normal random variable.

References

[Ahm12] A. Ahmed. Simplified and advanced approaches for the probabilistic analysis of shallow foundations. PhD thesis, University of Nantes, 2012.

[Au01] S.K. Au and J.L. Beck. Estimation of small failure probabilities in high dimentions by subset simulation, *Journal of Probabilistic Engineering Mechanics*, 16: 263-277, 2001.

[Hua09] H. Huang, B. Liang and KK Phoon. Geotechnical probabilistic analysis by collocation-based stochastic response surface method: An EXCEL add-in implementation. *Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards*, 3(2):75-86.

[Pho08] K.K. Phoon. *Reliability-based design in geotechnical engineering: Computations and applications*, Taylor & Francis, 1-75, 2008.

[San11] A.M. Santoso, K.K. Phoon and S.T. Quek. Modified Metropolis-Hastings algorithm with reduced chain-correlation for efficient subset simulation. *Probabilistic Engineering Mechanics*, 26(2): 331-341, 2011.

[Sud08] B. Sudret. Global sensitivity analysis using polynomial chaos expansion. *Reliability Engineering and System Safety,* 93: 964-979.

Random Fields

Gordon A. Fenton

Dalhousie University, Canada

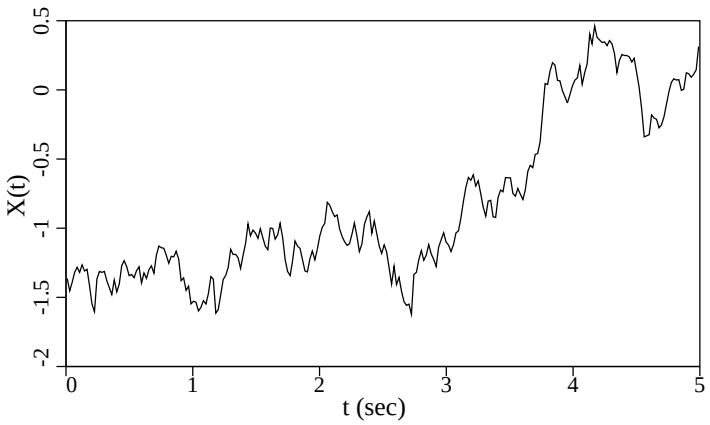
To model spatially variable geotechnical properties, we need to consider models more complex than the simple random variables we have seen so far in this course. We now need to allow every point in our ground model to become a random variable. In other words, we need to represent the ground using random fields, where a random field is defined to be an organized collection of ordinary random variables, one for each point in the ground. In general, random fields are three-dimensional, and although the fourth dimension of time can easily be added, at least conceptually, we will consider only spatial variability in this chapter. The chapter starts by discussing how random fields are characterized and how their basic statistical parameters are estimated. It then reviews how known information (sample data) can be used to best estimate the ground properties at a location which has not been sampled using Best Linear Unbiased Estimation or Kriging. Finally, the methods of random simulation of ground property random fields, so that probabilistic geotechnical questions can be answered, are discussed in detail.

1 Basic Random Field Concepts

Let's start by considering the simplest of random fields, that along a 1-D line in time or space. This could, for example, be the variation of a CPT sounding with depth, or it could be the ground acceleration at a particular location during an earthquake. We will use the earthquake time variation to illustrate the basic concepts of random field theory, but do keep in mind that these concepts are equally easily applied to space.

Consider the random field, or process, $X(t)$, which is made up of a sequence of random variables, $X(t)$, $X(t+dt)$, ..., taking on an infinite number of possible values on the real line.

Consider the following possible realization of $X(t)$.



The above figure could, for example, be the acceleration felt at a particular location in the ground during an earthquake. Although the above might have been measured, and so is a *realization*, we know that the accelerations felt at a neighboring location, or during a future earthquake (and even in the next few seconds) might be quite different than that recorded above. In general, we will not know the details of future ground motions and so we must assume them to be uncertain. They are therefore amenable to characterization by random fields, since the entire point of probability theory is to rationally characterize our uncertainty.

Since random fields consist of a collection of random variables, they are completely specified by the joint distribution between all of their component random variables. Since a random field is made up of an infinite number of spatial points, each of which has its own associated random variable, the complete specification of a random field would be an infinite-dimensional joint probability distribution. How do we characterize such a process? To do so, we should consider

1. *variability at a point:* pick an instant in time, t^*. At this point the process has random value $X(t^*) = X^*$ which is governed by some probability density function, $f_{X^*}(x)$. If we picked another point in time, say t', then $X(t') = X'$ would have another, possibly different PDF, $f_{X'}(x)$. That is, the PDF's could evolve with time (although this is complicated and hard to estimate in practice, in general, unless the PDF's are evolving merely by a simple trend in the mean or standard deviation).
   An example where the point, or marginal, distribution evolves with time is in earthquake ground motion where usually the variance increases drastically during the strong motion portion of the record.

2. *spatial dependence:* Consider again the two points in time, t^* and t'. If $X(t^*)$ and $X(t')$ are independent for any time lag $\tau = t' - t^*$, then the process would be infinitely rough – points separated by vanishingly small lags could have quite different values. This is not physically realistic for most natural phenomena. Thus, $X(t^*)$ and $X(t')$ generally have some sort of interdependence

(that often decreases with separation distance). This interdependence results in a smoothing of the random process. That is, for small τ, nearby states of X are preferential – the random field is constrained by its neighbors. We characterize the interdependence with the joint distribution $f_{X^*X'}(x^*, x')$. If we extend this idea to the consideration of any three, or four, or five, ..., points then the complete probabilistic description of a random process is the infinite-dimensional probability density function $f_{X_1X_2\ldots}(x_1, x_2, \ldots)$.

Of course, such an infinite-dimensional distribution would be difficult to fully describe, or more precisely, would only be reasonably properly described if an infinite number of realizations of the random field were available to use in estimating the parameters of the infinite-dimensional probability distribution. Usually, we have just one realization of the ground at a site, from which we must develop a reasonable model of our uncertainty regarding its spatial variability. From one realization, we cannot estimate the parameters of a general infinite-dimensional probability distribution, so we introduce a number of common simplifying assumptions to reduce the number of unknown parameters;

1. *Gaussian process:* the joint PDF is a multivariate normally distributed random process. The great advantage to the multivariate normal distribution is that the complete distribution can be specified by just the mean vector and the covariance matrix. The multivariate normal PDF has the form

$$f_{X_1X_2\ldots,X_k}(x_1, x_2, \ldots, x_k) = \frac{1}{(2\pi)^{k/2}} \frac{1}{|\underset{\approx}{C}|^{1/2}} \exp\left\{-\tfrac{1}{2}(\underset{\sim}{x} - \underset{\sim}{\mu})^T \underset{\approx}{C}^{-1}(\underset{\sim}{x} - \underset{\sim}{\mu})\right\} \tag{1}$$

where $\underset{\sim}{\mu}$ is the vector of mean values, one for each X_i, $\underset{\approx}{C}$ is the covariance matrix between the X's, and $|\underset{\approx}{C}|$ is its determinant. Specifically, $\underset{\sim}{\mu} = \mathrm{E}\big[\underset{\sim}{X}\big]$ and

$$\underset{\approx}{C} = \mathrm{E}\big[(\underset{\sim}{X} - \underset{\sim}{\mu})(\underset{\sim}{X} - \underset{\sim}{\mu})^T\big] \tag{2}$$

where the superscript T means the transpose. The covariance matrix, $\underset{\approx}{C}$, is a $k \times k$ symmetric, positive definite, matrix. For a continuous random field, the dimensions of $\underset{\sim}{\mu}$ and $\underset{\approx}{C}$ are still infinite, since the random field is composed of an infinite number of X's, one for each point. However, we often quantify $\underset{\sim}{\mu}$ and $\underset{\approx}{C}$ using continuous functions of space. For example, in a one-dimensional random field (or random process), the mean may vary linearly, i.e., $\mu(t) = a+bt$, and the covariance matrix can be expressed in terms of the standard deviations, which may vary with t, and the correlation function, ρ, as in

$$C(t_1, t_2) = \sigma(t_1)\sigma(t_2)\rho(t_1, t_2) \tag{3}$$

which specifies the covariance between $X(t_1)$ and $X(t_2)$. It the random field is stationary, then $\mu(t) = \mu$ is constant, Again, such a PDF is difficult to use in practice, not only mathematically, but also to estimate from real data.

2. *stationarity or statistical homogeneity:* the joint PDF is independent of spatial position, that is it just depends on relative positions of the points, not on their

position along the real line. This assumption implies that the mean, variance, and higher order moments are constant in time (or space) and thus that the marginal, or point, PDF is also constant in time (or space). So called *weak stationarity* or *second order stationarity* just implies that the mean and variance are constant in space.

3. *isotropy:* in two and higher dimensional random fields, isotropy implies that the joint PDF is invariant under rotation. This condition implies statistical homogeneity. For our purposes, isotropy usually just means that the correlation between two points only depends on the distance between the two points, not on their orientation relative to one another.

Some comments need to be made about the above. First of all, a random field, $X(t)$ having non-stationary mean and variance can be converted to a random field which is stationary in the mean and variance by the following transformation;

$$X'(t) = \frac{X(t) - \mu(t)}{\sigma(t)} \tag{4}$$

In fact, the random field $X'(t)$ will now have zero mean and unit variance everywhere. Similarly, a non-stationary random field can be produced from a stationary random field. For example, if $X(t)$ is a standard Gaussian random field (having zero mean and unit variance) and

$$Y(t) = 2 + \tfrac{1}{2}t + \tfrac{1}{4}\sqrt{t}X(t) \tag{5}$$

then $Y(t)$ is a non-stationary Gaussian random field with

$$\mathrm{E}\big[Y(t)\big] = \mu_Y(t) = 2 + \tfrac{1}{2}t \quad \text{and} \quad \mathrm{Var}\big[Y(t)\big] = \sigma_Y^2(t) = \tfrac{1}{2}t \tag{6}$$

in which both the mean and variance increase with t.

Secondly, we can often produce a non-Gaussian random field simply by transforming a Gaussian random field. For example, the random field $Y(t)$ defined by

$$Y(t) = e^{X(t)} \tag{7}$$

will have a lognormal distribution if $X(t)$ is a Gaussian process. A note of caution here, however, is that the covariance structure of the resulting field is also non-linearly transformed. For example, if $X(1)$ has correlation coefficient 0.2 with $X(2)$, the same is no longer true of $Y(1)$ and $Y(2)$. In fact, the correlation function of Y is now given by

$$\rho_Y(\tau) = \frac{\exp\{\sigma_X^2 \rho_X(\tau)\} - 1}{\exp\{\sigma_X^2\} - 1} \tag{8}$$

At this point, we can, in principle, describe a Gaussian random field and ask probabilistic questions of it.

One useful result relating to a Gaussian random field is that if one or more points in the field have been observed, the remainder of the field conditioned on those observations

remains Gaussian. For example, suppose $X(t)$ is observed to be 3.2 and one wonders what $X(t+1)$ is. If $X(t)$ and $X(t+1)$ are strongly correlated, then we would expect $X(t+1)$ to be close to 3.2 in value. Alternatively, if $X(t)$ and $X(t+1)$ are uncorrelated, then knowledge of $X(t)$ tells us nothing about $X(t+1)$. For simplicity, let us denote our observed value of $X(t)$ by simply x (i.e. $X(t) = x$), and let $Y = X(t+s)$ be some unobserved point. If X is a Gaussian process, then Y is normally distributed with conditional mean and variance given by

$$\mu_{Y \mid X} = \mu_Y + \rho(s)(x - \mu_X)\sigma_Y/\sigma_X \tag{9a}$$

$$\sigma_{Y \mid X} = \sigma_Y \sqrt{1 - \rho^2(s)} \tag{9b}$$

where $\rho(s)$ is the correlation coefficient between $X(t)$ and $X(t+s)$. If the random process is stationary, then $\mu_Y = \mu_X$ and $\sigma_Y = \sigma_X$ and the above conditional mean simplifies to $\mu_{Y \mid X} = \mu_X + \rho(s)(x - \mu_X)$.

1.1 The Variance Function

Virtually all engineering properties are actually properties of a local average of some sort. For example, hydraulic conductivity is generally obtained using a laboratory sample of some size, supplying a water pressure, and measuring the volume of water which passes through the sample in some time interval. The paths that the water takes to migrate through the sample are not considered individually, rather it is the sum of these paths that are measured. This is a 'local average' over the laboratory sample.

Similarly, when the compressive strength of a material is determined, a load is applied to a finite sized sample until failure occurs. Failure takes place when the shear/tensile resistance of a large number of bonds are broken – the failure load is a function of the average bond strength along the failure surface.

Thus, it is of considerable engineering interest to investigate how averages of random fields behave. Consider a local average defined as

$$X_T(t) = \frac{1}{T} \int_{t-T/2}^{t+T/2} X(\xi)\, d\xi \tag{10}$$

which is a 'moving' local average. That is, $X_T(t)$ is the local average of $X(t)$ under a window of width T centered at t. As this window is moved along in time, the local average $X_T(t)$ changes more slowly.

For example, consider the example of something floating on the ocean's surface: if the motion of a speck of sawdust on the surface of the ocean is tracked, it is seen to have considerable variability in its elevation. In fact, it will have as much variability as the waves themselves. Now, replace the sawdust with an ocean liner. The liner does not bounce around with every wave, but rather it 'averages' out the wave motion over the area of the liner. In other words, its vertical variability is drastically reduced.

In the above example, it is also worth thinking about the spectral representation of the ocean waves. The speck of sawdust sees all of the waves, big and small, whereas the local averaging taking place over the ocean liner damps out the high frequency components leaving just the long wavelength components (wavelengths of the order of the size of the ship and longer). Thus, local averaging is a low-pass filter. If the ocean waves on the day that the sawdust and ocean liner are being observed are composed of just long wavelength swells, then the variability of the sawdust and liner will be the same. Conversely, if the ocean surface is just choppy without any swells, then the ocean liner may not move up and down at all. However, both the sawdust and the ocean liner will have the same mean elevation in all cases.

So, the main effect of local averaging is to reduce the variance, and the amount of variance reduction increases with increasing high-frequency content in the random field. An increased high-frequency content corresponds to increasing independence in the random field, so that another way of putting this is that variance reduction increases when the random field consists of more 'independence'.

Let us look in more detail at the moments of $X_T(t)$. Its mean is

$$\mathrm{E}\big[X_T(t)\big] = \mathrm{E}\left[\frac{1}{T}\int_{t-T/2}^{t+T/2} X(\xi)\,d\xi\right] = \frac{1}{T}\int_{t-T/2}^{t+T/2} \mathrm{E}\big[X(\xi)\big]\;d\xi = \mathrm{E}\big[X\big] \tag{11}$$

for stationary $X(t)$. That is, local averaging preserves the mean of the random field (the mean of an arithmetic average is just the mean of the process). Consider the variance,

$$\mathrm{Var}\big[X_T(t)\big] = \mathrm{E}\big[(X_T(t) - \mu_{X_T})^2\big] \tag{12}$$

where, since $\mu_{X_T} = \mu_X$,

$$X_T - \mu_{X_T} = \frac{1}{T}\int_{t-T/2}^{t+T/2} X(\xi)\,d\xi - \mu_X = \frac{1}{T}\int_{t-T/2}^{t+T/2} [X(\xi) - \mu_X]\,d\xi \tag{13}$$

so that (due to stationarity, the bounds of the integral can be changed to any domain of length T without changing the expectation; we will use the domain $[0, T]$ for simplicity),

$$\begin{aligned}
\mathrm{Var}\big[X_T(t)\big] &= \mathrm{E}\left[\frac{1}{T}\int_0^T [X(\xi) - \mu_X]\,d\xi \cdot \frac{1}{T}\int_0^T [X(\eta) - \mu_X]\,d\eta\right] \\
&= \frac{1}{T^2}\int_0^T\int_0^T \mathrm{E}\big[(X(\xi) - \mu_X)(X(\eta) - \mu_X)\big]\;d\xi\,d\eta \\
&= \frac{1}{T^2}\int_0^T\int_0^T C_X(\xi - \eta)\,d\xi\,d\eta \;=\; \frac{\sigma_X^2}{T^2}\int_0^T\int_0^T \rho_X(\xi - \eta)\,d\xi\,d\eta \\
&= \sigma_X^2\,\gamma(T)
\end{aligned} \tag{14}$$

where $C_X(\tau)$ is the covariance function of $X(t)$, and $\rho_X(\tau)$ is the correlation function of $X(t)$, such that $C_X(\tau) = \sigma_X^2 \rho_X(\tau)$. In the final expression, $\gamma(T)$ is the so-called

variance function, which gives the amount that the variance is reduced when averaged over the length T. The variance function has value 1.0 when $T = 0$, which is to say that $X_T(t) = X(t)$ when $T = 0$ and so the variance is not at all reduced. As T increases, the variance function falls towards zero. It has the mathematical definition

$$\gamma(T) = \frac{1}{T^2} \int_0^T \int_0^T \rho_X(\xi - \eta)\, d\xi\, d\eta \tag{15}$$

If one considers this integral which is over the square region $[0, T] \times [0, T]$ in (ξ, η) space, one sees that $\rho_X(\xi - \eta)$ is constant along diagonal lines where $\xi - \eta =$ constant. The length of the main diagonal, where $\xi = \eta$, is $\sqrt{2}T$, and the other diagonal lines decreasing linearly in length to zero in the corners. Thus, the double integral can be collapsed to a single integral by integrating in a direction perpendicular to the diagonals; each diagonal line has length $\sqrt{2}(T - |\tau|)$, width $d\tau/\sqrt{2}$, and 'height' equal to $\rho_X(\xi - \eta) = \rho_X(\tau)$ The integral reduces to

$$\gamma(T) = \frac{1}{T^2} \int_{-T}^{T} \sqrt{2}(T - |\tau|)\rho_X(\tau)\, \frac{d\tau}{\sqrt{2}}] = \frac{1}{T^2} \int_{-T}^{T} (T - |\tau|)\rho_X(\tau)\, d\tau \tag{16}$$

Furthermore, since $\rho_X(\tau) = \rho_X(-\tau)$, the integrand is even giving us

$$\gamma(T) = \frac{2}{T^2} \int_0^T (T - \tau)\rho_X(\tau)\, d\tau \tag{17}$$

The variance function can be seen in Eq. (15) above to be an average of the correlation coefficient between every pair of points on the interval $[0, T]$. If the correlation function falls off rapidly, so that the correlation between pairs of points becomes rapidly smaller with separation distance, then $\gamma(T)$ will be small. On the other hand, if all points on the interval $[0, T]$ are perfectly correlated, having $\rho(\tau) = 1$ for all τ, then $\gamma(T)$ will be 1.0. Such a field displays no variance reduction under local averaging. (In fact, in such a field all points have the same random value, $X(t) = X$, if the field is stationary.)

The variance function is another 'equivalent' second-moment description of a random field, since it can be obtained through knowledge of the correlation function, which in turn can be obtained from the spectral density function. The inverse relationship between $\gamma(T)$ and $\rho(\tau)$ is obtained by differentiation;

$$\rho(\tau) = \frac{1}{2} \frac{d^2}{d\tau^2} [\tau^2 \gamma(\tau)] \tag{18}$$

1.2 The Scale of Fluctuation

A convenient measure of the variability of a random field is the so-called *scale of fluctuation*, θ. Loosely speaking θ is the distance beyond which points are largely

uncorrelated. Conversely, two points separated by a distance less than θ will be significantly correlated. Mathematically, θ can be defined as the area under the correlation function;

$$\theta = \int_{-\infty}^{\infty} \rho(\tau)\, d\tau = 2\int_{0}^{\infty} \rho(\tau)\, d\tau \tag{19}$$

This relationship implies that if θ is to exist (ie. be finite) that $\rho(\tau)$ must decrease sufficiently quickly to zero as τ increases. Not all valid correlation functions will satisfy this criteria so that for such random processes, $\theta = \infty$. An example of a process with infinite scale of fluctuation is a *fractal* or statistically self-similar process.

In addition, the scale of fluctuation is really only meaningful for strictly non-negative correlation functions. Since $-1 \leq \rho \leq 1$, one could conceivably have an oscillatory correlation function whose area is zero but which has significant correlations (positive or negative) over significant distances. An example of such a correlation function might be that governing wave heights in a body of water. We will not consider such cases since most engineering materials have strictly non-negative correlation functions (an exception possibly being laminates).

The scale of fluctuation can also be defined in terms of the spectral density function (see next Section),

$$G(\omega) = \frac{2\sigma^2}{\pi}\int_{0}^{\infty} \rho(\tau)\cos(\omega\tau)\, d\tau \tag{20}$$

so that

$$G(0) = \frac{2\sigma^2}{\pi}\int_{0}^{\infty} \rho(\tau)\, d\tau = \frac{\sigma^2}{\pi}\theta \tag{21}$$

or

$$\theta = \frac{\pi G(0)}{\sigma^2} \tag{22}$$

which means that if the spectral density function is finite at the origin, then θ will exist.

Finally, the scale of fluctuation can also be defined in terms of the variance function as a limit;

$$\theta = \lim_{T\to\infty} T\gamma(T) \tag{23}$$

In turn, this implies that if the scale of fluctuation is finite, that the variance function has the following limiting form as the averaging region grows very large;

$$\lim_{T\to\infty} \gamma(T) = \frac{\theta}{T} \tag{24}$$

which in turn means that θ/T can be used as an approximation for $\gamma(T)$ when $T >> \theta$. A more extensive approximation for $\gamma(T)$, useful when the precise correlation structure of a random field is unknown, but for which θ is known (or estimated) is

$$\gamma(T) \simeq \frac{\theta}{\theta + |T|} \tag{25}$$

which has the correct limiting form for $T >> \theta$ and which has value 1.0 when $T = 0$, as expected.

Several commonly used correlation functions are parameterized by the scale of fluctuation, for example,

1. *Markov correlation function:*

$$\rho(\tau) = \exp\left\{-\frac{2|\tau|}{\theta}\right\} \tag{26}$$

which has variance function

$$\gamma(T) = \frac{\theta^2}{2T^2}\left[\frac{2|T|}{\theta} + \exp\left\{-\frac{2|T|}{\theta}\right\} - 1\right] \tag{27}$$

and spectral density function

$$G(\omega) = \frac{4\sigma^2\theta}{\pi(4+\theta^2\omega^2)} \tag{28}$$

2. *Gaussian type correlation function:*

$$\rho(\tau) = \exp\left\{-\pi\left(\frac{|\tau|}{\theta}\right)^2\right\} \tag{29}$$

which has variance function

$$\gamma(T) = \frac{\theta^2}{\pi T^2}\left[\frac{\pi|T|}{\theta}\mathrm{erf}\left\{\frac{\sqrt{\pi}|T|}{\theta}\right\} + \exp\left\{-\frac{\pi T^2}{\theta^2}\right\} - 1\right] \tag{30}$$

where $\mathrm{erf}(x) = 2\Phi(\sqrt{2}x) - 1$ is the error function and $\Phi(z)$ is the standard normal cumulative distribution function.

3. *Fractional Gaussian Noise:* this is a form of a fractal process as defined by Mandelbrot and Ness,

$$\rho(\tau) = \frac{1}{2\delta^{2H}}\left[|\tau+\delta|^{2H} - 2|\tau|^{2H} + |\tau-\delta|^{2H}\right], \tag{31a}$$

$$\gamma(T) = \frac{|T+\delta|^{2H+2} - 2|T|^{2H+2} + |T-\delta|^{2H+2} - 2\delta^{2H+2}}{T^2(2H+1)(2H+2)\delta^{2H}}, \tag{31b}$$

defined for $0 < H < 1$. The case $H = 0.5$ corresponds to white noise. Note that this process actually has two parameters, H and δ. The latter is a small averaging region thrown in to damp out the infinite variance high frequency contributions to the true fractal process – that is, it changes the infinite variance fractal process into a finite variance 'band-limited' approximation to the fractal process.

4. *Polynomial decaying correlation function:*

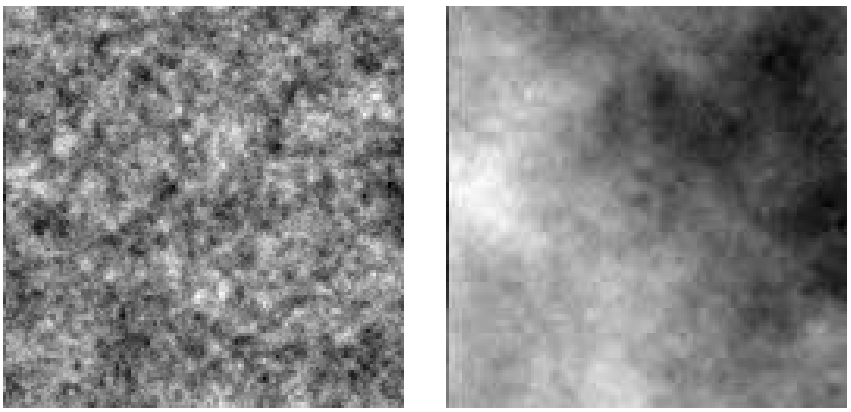
$$\rho(\tau) = \frac{\theta^3}{(\theta + \tau)^3} \tag{32}$$

which has variance function

$$\gamma(T) = \frac{\theta}{\theta + T} \tag{33}$$

Finally, some comments about what affect the scale of fluctuation has on a random field are in order. When the scale of fluctuation is small, the field tends to be somewhat 'rough'. In the limit, for say a Markov correlation function, the field becomes a so-called white noise when $\theta \to 0$. Conversely, when the scale of fluctuation is large, the field is smoother. In the limit, for say a Markov correlation function, the field becomes completely uniform – different from realization to realization but each realization is composed of a single random value.

The next figure shows two random field realizations, each of size 1×1. The field on the left has a small scale of fluctuation ($\theta = 0.04$) and can be seen to be quite rough. The field on the right has a large scale of fluctuation ($\theta = 2$) and can be seen to be more slowly varying.



1.3 The Spectral Density Function

We now turn our attention to an equivalent 2nd-order description of a stationary random process, namely its *spectral representation*. We say 'equivalent' because the spectral representation, in the form of a *spectral density function*, contains the same information as the covariance function, just expressed in a different way. As we shall see, the spectral density function can be obtained from the covariance function and vice-versa. The two forms are merely transforms of one another.

Priestley (1981) shows that if $X(t)$ is a stationary random process, with $\rho(\tau)$ continuous at $\tau = 0$, then it can be expressed as a sum of sinusoids with mutually independent random amplitudes and phase angles,

$$X(t) = \mu_X + \sum_{k=-N}^{N} C_k \cos(\omega_k t + \Phi_k) = \mu_X + \sum_{k=-N}^{N} \Big(A_k \cos(\omega_k t) + B_k \sin(\omega_k t) \Big) \tag{34}$$

where μ_X is the process mean, C_k is a random amplitude, and Φ_k is a random phase angle. The equivalent form involving A_k and B_k is obtained by setting $A_k = C_k \cos(\Phi_k)$ and $B_k = -C_k \sin(\Phi_k)$. If the random amplitudes A_k and B_k are normally distributed with zero means, then $X(t)$ will also be normally distributed with mean μ_X. For this to be true, C_k must be Raleigh distributed and Φ_k must be uniformly distributed on the interval $[0, 2\pi]$. Note that $X(t)$ will tend to a normal distribution anyhow, by virtue of the central limit theorem, for wide-band processes, so we will assume that $X(t)$ is normally distributed.

Consider the k^{th} component of $X(t)$, and ignore μ_X for the time being,

$$X_k(t) = C_k \cos(\omega_k t + \Phi_k) \tag{35}$$

If C_k is independent of Φ_k, then $X_k(t)$ has mean

$$\mathrm{E}\big[X_k(t)\big] = \mathrm{E}\big[C_k \cos(\omega_k t + \Phi_k)\big] = \mathrm{E}\big[C_k\big]\,\mathrm{E}\big[\cos(\omega_k t + \Phi_k)\big] = 0 \tag{36}$$

due to independence and the fact that for any t, $\mathrm{E}\big[\cos(\omega_k t + \Phi_k)\big] = 0$ since Φ_k is uniformly distributed on $[0, 2\pi]$. The variance of $X_k(t)$ is thus

$$\mathrm{Var}\big[X_k(t)\big] = \mathrm{E}\big[X_k^2(t)\big] = \mathrm{E}\big[C_k^2\big]\,\mathrm{E}\big[\cos^2(\omega_k t + \Phi_k)\big] = \tfrac{1}{2}\mathrm{E}\big[C_k^2\big] \tag{37}$$

Note that $\mathrm{E}\big[\cos^2(\omega_k t + \Phi_k)\big] = \frac{1}{2}$, which again uses the fact that Φ_k is uniformly distributed between 0 and 2π.

Priestley also shows that the component sinusoids are independent of one another, that is that $X_k(t)$ is independent of $X_j(t)$, for all $k \neq j$. Using this property, we can put the components back together to find the mean and variance of $X(t)$,

$$\mathrm{E}\big[X(t)\big] = \mu_X + \sum_{k=-N}^{N} \mathrm{E}\big[X_k(t)\big] = \mu_X \tag{38a}$$

$$\mathrm{Var}\big[X(t)\big] = \sum_{k=-N}^{N} \mathrm{Var}\big[X_k(t)\big] = \sum_{k=-N}^{N} \tfrac{1}{2}\mathrm{E}\big[C_k^2\big] \tag{38b}$$

In other words, the prescribed mean of $X(t)$ is preserved by the spectral representation and the variance of the sum is the sum of the variances of each component frequency, since the component sinusoids are independent. The amount that each component frequency contributes to the overall variance of $X(t)$ depends on the 'power' in the sinusoid amplitude, $\frac{1}{2}\mathrm{E}\big[C_k^2\big]$.

Now define the *two-sided spectral density function*, $S(\omega)$, such that

$$S(\omega_k)\Delta\omega = \mathrm{Var}\big[X_k(t)\big] = \mathrm{E}\big[X_k^2(t)\big] = \tfrac{1}{2}\mathrm{E}\big[C_k^2\big] \tag{39}$$

then the variance of $X(t)$ can be written as

$$\mathrm{Var}\big[X(t)\big] = \sum_{k=-N}^{N} S(\omega_k)\Delta\omega \tag{40}$$

In the limit as $\Delta\omega \to 0$ and $N \to \infty$, we get

$$\mathrm{Var}\big[X(t)\big] = \sigma_X^2 = \int_{-\infty}^{\infty} S(\omega)\, d\omega \tag{41}$$

which is to say, the variance of $X(t)$ is just the area under the two-sided spectral density function.

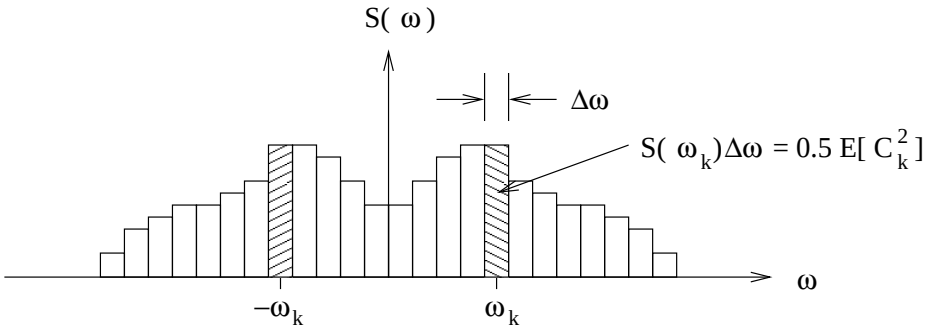


Figure 1. Two-sided spectral density function, $S(\omega)$.

1.3.1 Wiener-Khinchine Relations

We can use the spectral representation to express the covariance function, $C(\tau)$. Assuming that $\mu_X = 0$ for the time being to simplify the algebra (this is not a restriction, the end results are the same even if $\mu_X \neq 0$), we have

$$\begin{aligned} C(\tau) &= \mathrm{Cov}\big[X(0), X(\tau)\big], \qquad \text{(due to stationarity)} \\ &= \mathrm{E}\left[\sum_k X_k(0) \sum_j X_j(\tau)\right] \\ &= \sum_k \sum_j \mathrm{E}\big[X_k(0)X_j(\tau)\big] \\ &= \sum_k \mathrm{E}\big[X_k(0)X_k(\tau)\big], \qquad \text{(due to independence)} \end{aligned} \tag{42}$$

Now, since $X_k(0) = C_k \cos(\Phi_k)$ and $X_k(\tau) = C_k \cos(\omega_k \tau + \Phi_k)$ we get

$$\begin{aligned} C(\tau) &= \sum_k \mathrm{E}\big[C_k^2\big]\, \mathrm{E}\big[\cos(\Phi_k)\cos(\omega_k\tau + \Phi_k)\big] \\ &= \sum_k \mathrm{E}\big[C_k^2\big]\, \mathrm{E}\big[\tfrac{1}{2}\{\cos(\omega_k\tau + 2\Phi_k) + \cos(\omega_k\tau)\}\big] \\ &= \sum_k \tfrac{1}{2}\mathrm{E}\big[C_k^2\big]\cos(\omega_k\tau) \\ &= \sum_k S(\omega_k)\cos(\omega_k\tau)\Delta\omega \end{aligned} \tag{43}$$

which, in the limit as $\Delta\omega \to 0$ gives

$$C(\tau) = \int_{-\infty}^{\infty} S(\omega)\cos(\omega\tau)\, d\omega \tag{44}$$

Thus, the covariance function $C(\tau)$ is the Fourier transform of the spectral density function, $S(\omega)$. The inverse transform can be applied to find $S(\omega)$ in terms of $C(\tau)$,

$$S(\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} C(\tau)\cos(\omega\tau)\, d\tau \tag{45}$$

so that knowing either $C(\tau)$ or $S(\omega)$ allows the other to be found (and hence these are 'equivalent' in terms of information). Also, since $C(\tau) = C(-\tau)$, ie. that the covariance between one point and another is the same regardless of which point you consider first, and since $\cos(x) = \cos(-x)$, we see that

$$S(\omega) = S(-\omega) \tag{46}$$

In other words, the two-sided spectral density function is an even function (see Figure 1). The fact that $S(\omega)$ is symmetric about $\omega = 0$ means that we need only know the positive half in order to know the entire function. This motivates the introduction of the *one-sided spectral density function*, $G(\omega)$ defined as

$$G(\omega) = 2S(\omega), \quad \omega \geq 0 \tag{47}$$

The factor of two is included to preserve the total variance when only positive frequencies are considered. Now the Wiener-Khinchine relations become

$$C(\tau) = \int_0^{\infty} G(\omega)\cos(\omega\tau)\, d\omega \tag{48a}$$

$$G(\omega) = \frac{1}{\pi}\int_{-\infty}^{\infty} C(\tau)\cos(\omega\tau)\, d\tau \tag{48b}$$

$$= \frac{2}{\pi}\int_0^{\infty} C(\tau)\cos(\omega\tau)\, d\tau \tag{48c}$$

and the variance of $X(t)$ is the area under $G(\omega)$ (set $\tau = 0$ in Eq. to see this),

$$\sigma_x^2 = C(0) = \int_0^\infty G(\omega)\, d\omega \tag{49}$$

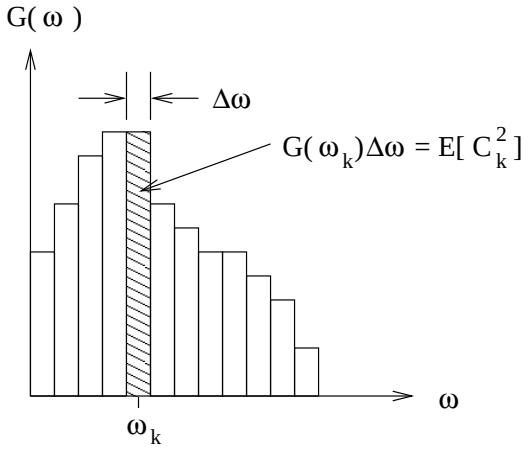


Figure 2. One-sided spectral density function, $G(\omega) = 2S(\omega)$ corresponding to Figure 1.

Some commonly used spectral density functions are as follows;

1. *White Noise:* the SDF has equal power at all frequencies (hence analogy with white light. The corresponding covariance structure is a delta function – all points are uncorrelated, which is simple. Unfortunately, the variance of white noise is infinite. In practice a finite variance band-limited form is usually used, namely,

$$G(\omega) = \begin{cases} G_o & \text{for } \omega_{min} \le \omega \le \omega_{max} \\ 0 & \text{otherwise} \end{cases} \tag{50}$$

   where G_o is a constant refered to as the *white noise intensity*. The variance of this process is $G_o \times (\omega_{max} - \omega_{min})$

2. *Fractal Noise:* here the SDF varies inversely with frequency,

$$G(\omega) = \frac{G_o}{\omega^\gamma} \tag{51}$$

   This is also called a *statistically self-similar* process. It is believed to characterize a large number of natural phenomenon, but also has infinite variance.

The spectral representation of a stationary Gaussian process is primarily used in situations where the frequency domain is an integral part of the problem being considered. For example, earthquake ground motions are often represented using the spectral density function because the motions are largely sinusoidal with frequency content dictated by resonance in the soil or rock through which the earthquake waves are traveling. In addition, the response of structures to earthquake motion is often performed

using Fourier response 'modes', each having its own resonance frequency. Thus, if a structure has a 1 Hz primary response mode (single mass-and-spring oscillation), then it is of interest to see what 'power' the input ground motion has at 1 Hz. This is given by $G(\omega_k)\Delta\omega$ at $\omega_k = 1$ Hz.

In addition, the spectral representation provides a means to simulate a stationary Gaussian process, namely to simulate independent realizations of C_k and Φ_k, for $k = 0, 1, \ldots, N$, and then recombine using the spectral representation. We shall see more of this in the paper on Simulation.

2 References

Priestley, M.B. (1981). *Spectral Analysis and Time Series*, Vol. 1: Univariate Series, Academic Press, New York, NY.

Best Linear Unbiased Estimation

Gordon A. Fenton

Dalhousie University, Canada

1 Introduction

We often want some way of best estimating 'future' events given past observations. For example, suppose that we have observed $X_1, X_2, \ldots, X_n$ and we want to estimate the optimal value for X_{n+1} using this information. One possibility is to write our estimate for X_{n+1} as a linear combination of our observations;

$$\hat{X}_{n+1} = \mu_{n+1} + \sum_{k=1}^{n} \beta_k (X_k - \mu_k) \tag{1}$$

where the hat indicates that this is an estimate of X_{n+1} and μ_k is the mean of X_k. Note that we need to know the means in order to form this estimate. Equation (1) is refered to as the Best Linear Unbiased Estimator (BLUE) for reasons we shall soon see.

The question now is what is the optimal vector of coefficients, $\underset{\sim}{\beta}$? We can define 'optimal' to be that which produces the minimum expected error between our estimate $\hat{X}_{n+1}$ and the true (but unknown) X_{n+1}. The *estimator error*, E, is given by

$$E = X_{n+1} - \hat{X}_{n+1} = X_{n+1} - \mu_{n+1} - \sum_{k=1}^{n} \beta_k (X_k - \mu_k) \tag{2}$$

To make this error as small as possible, its mean should be zero and its variance minimal. The first criteria is automatically satisfied by the above formulation since

$$\begin{aligned} \mathrm{E}\big[E\big] = \mathrm{E}\Big[X_{n+1} - \hat{X}_{n+1}\Big] &= \mathrm{E}\left[X_{n+1} - \mu_{n+1} - \sum_{k=1}^{n} \beta_k (X_k - \mu_k)\right] \\ &= \mu_{n+1} - \mu_{n+1} - \sum_{k=1}^{n} \beta_k \mathrm{E}\big[X_k - \mu_k\big] \;=\; 0 \end{aligned} \tag{3}$$

We say that the estimator $\hat{X}_{n+1}$ is *unbiased* because its mean is the same as that being estimated, i.e., $\mathrm{E}\Big[\hat{X}_{n+1}\Big] = \mathrm{E}\big[X_{n+1}\big]$ by virtue of the above result.

Now we want to minimize the error variance. Since the mean estimator error is zero, its variance is just the expectation of the squared estimator error,

$$\mathrm{Var}\Big[X_{n+1} - \hat{X}_{n+1}\Big] = \mathrm{E}\Big[\Big(X_{n+1} - \hat{X}_{n+1}\Big)^2\Big] = \mathrm{E}\Big[X_{n+1}^2 - 2X_{n+1}\hat{X}_{n+1} + \hat{X}_{n+1}^2\Big] \tag{4}$$

To simplify the following algebra, we will assume that $\mu_i = 0$ for $i = 1, 2, \ldots, n+1$. The final result will be the same even if the means are non-zero, since variances and covariances are mean centered. In this case, our estimator simplifies to

$$\hat{X}_{n+1} = \sum_{k=1}^{n} \beta_k X_k \tag{5}$$

and the estimator error variance becomes

$$\mathrm{Var}\big[E\big] = \mathrm{E}\big[X_{n+1}^2\big] - 2\sum_{k=1}^{n} \beta_k \mathrm{E}\big[X_{n+1}X_k\big] + \sum_{k=1}^{n}\sum_{j=1}^{n} \beta_k\beta_j \mathrm{E}\big[X_k X_j\big] \tag{6}$$

To minimize this with respect to our unknown coefficients, $\beta_1, \beta_2, \ldots, \beta_n$, we set the derivative of the error to zero with respect to each unknown,

$$\frac{\partial}{\partial \beta_\ell}\mathrm{Var}\big[E\big] = 0 \qquad \text{for } \ell = 1, 2, \ldots, n \tag{7}$$

which gives us n equations in n unknowns. Differentiating each term in Eq. (6) leads to,

$$\begin{aligned}
\frac{\partial}{\partial \beta_\ell}\mathrm{E}\big[X_{n+1}^2\big] &= 0 \\
\frac{\partial}{\partial \beta_\ell}\sum_{k=1}^{n} \beta_k \mathrm{E}\big[X_{n+1}X_k\big] &= \mathrm{E}\big[X_{n+1}X_\ell\big] \\
\frac{\partial}{\partial \beta_\ell}\sum_{k=1}^{n}\sum_{j=1}^{n} \beta_k\beta_j \mathrm{E}\big[X_k X_j\big] &= 2\sum_{k=1}^{n} \beta_k \mathrm{E}\big[X_\ell X_k\big]
\end{aligned} \tag{8}$$

which gives us

$$\frac{\partial}{\partial \beta_\ell}\mathrm{Var}\big[E\big] = -2\mathrm{E}\big[X_{n+1}X_\ell\big] + 2\sum_{k=1}^{n} \beta_k \mathrm{E}\big[X_\ell X_k\big] = 0 \tag{9}$$

This means that

$$\mathrm{E}\big[X_{n+1}X_\ell\big] = \sum_{k=1}^{n} \beta_k \mathrm{E}\big[X_\ell X_k\big] \tag{10}$$

for $\ell = 1, 2, \ldots, n$. If we define the following matrix and vector components

$$C_{\ell k} = \mathrm{E}\big[X_\ell X_k\big] = \mathrm{Cov}\big[X_\ell, X_k\big] \tag{11a}$$
$$b_\ell = \mathrm{E}\big[X_\ell X_{n+1}\big] = \mathrm{Cov}\big[X_\ell, X_{n+1}\big] \tag{11b}$$

then Eq. (10) can be written as

$$b_\ell = \sum_{k=1}^{n} C_{\ell k}\beta_k \tag{12}$$

or, in matrix notation

$$\underset{\sim}{b} = \underset{\approx}{C}\underset{\sim}{\beta} \tag{13}$$

which has solution

$$\beta = \underset{\approx}{C}^{-1}\underset{\sim}{b} \tag{14}$$

There is some similarity between the above Best Linear Unbiased Estimate and regression analysis. The primary difference is that regression ignors correlations between data points and considers only distance between data points (along with the value at the data point, of course). BLUE replaces 'distance' with covariance, but requires that both the mean and covariance are known ahead of time. We shall see in the next Section that Kriging relieves us of having to know the mean ahead of time so that 'distance' in regression is replaced only by knowledge of the covariance.

Example 1:

Suppose that gravity measurements at a site suggests that the mean gold concentration along a seam, in parts-per-million, shows a slow increase with the distance s, in metres, along the seam, that is that

$$\mu(s) = 2000 + 300s \tag{15}$$

Furthermore suppose that a statistical analysis of a similar site has given the following covariance function which is assumed to also hold at the current site,

$$C(\tau) = \sigma_X^2 \exp\left\{-\frac{|\tau|}{4}\right\} \tag{16}$$

where $\sigma_X = 500$ ppm and where τ is the separation distance between points. We want to estimate the gold concentration at $s = 3$ m, given the following observations at $s = 1$ and $s = 2$

$$\text{at } s = 1\text{:} \quad x_1 = 2130 \quad \textit{ppm}$$
$$\text{at } s = 2\text{:} \quad x_2 = 2320 \quad \textit{ppm}$$

Solution:

We start by finding the components of the covariance matrix and vector;

$$\underset{\sim}{b} = \left\{ \begin{array}{c} \mathrm{Cov}\big[X_1, X_3\big] \\ \mathrm{Cov}\big[X_2, X_3\big] \end{array} \right\} = \sigma_x^2 \left\{ \begin{array}{c} e^{-2/4} \\ e^{-1/4} \end{array} \right\} \tag{17}$$

$$\underset{\approx}{C} = \left[\begin{array}{cc} \mathrm{Cov}\big[X_1, X_1\big] & \mathrm{Cov}\big[X_1, X_2\big] \\ \mathrm{Cov}\big[X_2, X_1\big] & \mathrm{Cov}\big[X_2, X_2\big] \end{array} \right] = \sigma_x^2 \left[\begin{array}{cc} 1 & e^{-1/4} \\ e^{-1/4} & 1 \end{array} \right] \tag{18}$$

Substituting these into Eq. (13) gives

$$\sigma_x^2 \left[\begin{array}{cc} 1 & e^{-1/4} \\ e^{-1/4} & 1 \end{array} \right] \left\{ \begin{array}{c} \beta_1 \\ \beta_2 \end{array} \right\} = \sigma_x^2 \left\{ \begin{array}{c} e^{-2/4} \\ e^{-1/4} \end{array} \right\} \tag{19}$$

Notice that the variance cancels out, which is typical when the variance is constant with position. We now get

$$\left\{ \begin{array}{c} \beta_1 \\ \beta_2 \end{array} \right\} = \left[\begin{array}{cc} 1 & e^{-1/4} \\ e^{-1/4} & 1 \end{array} \right]^{-1} \left\{ \begin{array}{c} e^{-2/4} \\ e^{-1/4} \end{array} \right\} = \left\{ \begin{array}{c} 0 \\ e^{-1/4} \end{array} \right\} \tag{20}$$

Thus, the optimal linear estimate of X_3 is

$$\begin{aligned} \hat{x}_3 &= \mu(3) + e^{-1/4}(x_2 - \mu(2)) = 2900 + e^{-1/4}(2320 - 2000 - 300(2)) \\ &= 2900 - 280e^{-1/4} \;=\; 2681 \quad \textit{ppm}(21) \end{aligned}$$

Notice that, because of the Markovian nature of the covariance function used in this example, the prediction of the 'future' depends only on the most recent 'past'. The prediction is independent of observations further in the 'past'. This is typical of the Markov correlation function in one-dimension (in higher dimensions, it is not quite so simple).

2 Estimator Error

Once the best linear unbiased estimate has been determined, it is of interest to ask how confident are we in this estimate? Can we assess the variability of our estimator? To investigate this, let us again consider a zero mean process so that our estimator can be simply written as

$$\hat{X}_{n+1} = \sum_{k=1}^{n} \beta_k X_k \tag{22}$$

In this case, the variance is simply determined as

$$\mathrm{Var}\Big[\hat{X}_{n+1}\Big] = \mathrm{Var}\left[\sum_{k=1}^{n} \beta_k X_k\right] \;=\; \mathrm{Var}\big[\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n\big] \tag{23}$$

The variance of a sum is the sum of the variances *only if the terms are independent.* In this case, the X's are not independent, so the *variance of a sum becomes the sum of all possible covariances*,

$$\mathrm{Var}\Big[\hat{X}_{n+1}\Big] = \sigma^2_{\hat{X}} = \sum_{k=1}^{n}\sum_{j=1}^{n}\beta_k\beta_j\mathrm{Cov}\big[X_k, X_j\big] = \underset{\sim}{\beta}^T \underset{\approx}{C}\underset{\sim}{\beta} \tag{24}$$

where T means transpose.

However, the above estimator variance is often of limited interest. We are typically more interested in asking questions such as: What is the probability that the true value of X_{n+1} exceeds our estimate, $\hat{X}_{n+1}$, by a certain amount. For example, we may want to compute

$$\mathrm{P}\Big[X_{n+1} > \hat{X}_{n+1} + b\Big] = \mathrm{P}\Big[X_{n+1} - \hat{X}_{n+1} > b\Big] \tag{25}$$

where b is some constant. Evidently, this would involve finding the distribution of the estimator error $E = (X_{n+1} - \hat{X}_{n+1})$. The variance of the estimator error can be found from Eq. (6) as follows,

$$\begin{aligned}
\sigma^2_{E} &= \mathrm{E}\big[X^2_{n+1}\big] - 2\sum_{k=1}^{n}\beta_k\mathrm{E}\big[X_{n+1}X_k\big] + \sum_{k=1}^{n}\sum_{j=1}^{n}\beta_k\beta_j\mathrm{E}\big[X_kX_j\big] \\
&= \sigma^2_X + \underset{\sim}{\beta}^T\underset{\approx}{C}\underset{\sim}{\beta} - 2\underset{\sim}{\beta}^T\underset{\sim}{b} \qquad \text{(rearranging terms)} \\
&= \sigma^2_X + \sigma^2_{\hat{X}} - 2\underset{\sim}{\beta}^T\underset{\sim}{b}
\end{aligned} \tag{26}$$

So we see that the variance of the estimator error (often refered to directly as the *estimator error*) is the sum of the variance in X and the variance in $\hat{X}$ less a term which depends on the degree of correlation between X and the observations. As the correlation between the observations and the point being estimated increases, it becomes less and less likely that the true value of X_{n+1} will stray very far from its estimate. So for high correlations between the observations and the estimated point, the estimator error becomes small. This can be seen more clearly if we simplify the estimator error equation. To do this, we note that $\underset{\sim}{\beta}$ has been determined such that $\underset{\approx}{C}\underset{\sim}{\beta} = \underset{\sim}{b}$, or, putting it another way, $\underset{\approx}{C}\underset{\sim}{\beta} - \underset{\sim}{b} = \underset{\sim}{0}$ (where $\underset{\sim}{0}$ is a vector of zeroes). Now we write

$$\begin{aligned}
\sigma^2_{E} &= \sigma^2_X + \underset{\sim}{\beta}^T\underset{\approx}{C}\underset{\sim}{\beta} - 2\underset{\sim}{\beta}^T\underset{\sim}{b} \\
&= \sigma^2_X + \underset{\sim}{\beta}^T\underset{\approx}{C}\underset{\sim}{\beta} - \underset{\sim}{\beta}^T\underset{\sim}{b} - \underset{\sim}{\beta}^T\underset{\sim}{b} \\
&= \sigma^2_X + \underset{\sim}{\beta}^T(\underset{\approx}{C}\underset{\sim}{\beta} - \underset{\sim}{b}) - \underset{\sim}{\beta}^T\underset{\sim}{b} \\
&= \sigma^2_X - \underset{\sim}{\beta}^T\underset{\sim}{b}
\end{aligned} \tag{27}$$

which is a much simpler way of computing σ^2_{E} and more clearly demonstrates the variance reduction due to correlation with observations.

The estimator $\hat{X}_{n+1}$ is also the conditional mean of X_{n+1} given the observations. That is

$$\mathrm{E}\big[X_{n+1} \,|\, X_1, X_2, \ldots, X_n\big] = \hat{X}_{n+1} \tag{28}$$

The conditional variance of X_{n+1} is σ_E^2,

$$\mathrm{Var}\big[X_{n+1} \,|\, X_1, X_2, \ldots, X_n\big] = \sigma_E^2 \tag{29}$$

Generally questions regarding the probability that the true X_{n+1} lies in some region should employ the conditional mean and variance of X_{n+1}, since this would then make use of all of the information at hand.

Example 2:

Consider again Example 1. What is the variance of the estimator and the estimator error? Estimate the probability that X_3 exceeds $\hat{X}_3$ by more than 400 ppm.

Solution:

We had

$$\underset{\approx}{C} = \sigma_X^2 \begin{bmatrix} 1 & e^{-1/4} \\ e^{-1/4} & 1 \end{bmatrix} = (500)^2 \begin{bmatrix} 1 & e^{-1/4} \\ e^{-1/4} & 1 \end{bmatrix} \tag{30}$$

and

$$\underset{\sim}{\beta} = \begin{Bmatrix} 0 \\ e^{-1/4} \end{Bmatrix} \tag{31}$$

so that

$$\sigma_{\hat{X}}^2 = \mathrm{Var}\Big[\hat{X}_3\Big] = (500)^2 \begin{Bmatrix} 0 & e^{-1/4} \end{Bmatrix} \begin{bmatrix} 1 & e^{-1/4} \\ e^{-1/4} & 1 \end{bmatrix} \begin{Bmatrix} 0 \\ e^{-1/4} \end{Bmatrix} = (500)^2\, e^{-2/4} \tag{32}$$

which gives $\sigma_{\hat{X}} = 500e^{-1/4} = 389.4$ ppm.

For the covariance vector found in Example 1,

$$\underset{\sim}{b} = \sigma_X^2 \begin{Bmatrix} e^{-2/4} \\ e^{-1/4} \end{Bmatrix} \tag{33}$$

the estimator error is computed as

$$\begin{aligned} \sigma_E^2 = \mathrm{Var}\Big[X_3 - \hat{X}_3\Big] &= \sigma_X^2 + \underset{\sim}{\beta}^T \underset{\approx}{C} \underset{\sim}{\beta} - 2\underset{\sim}{\beta}^T \underset{\sim}{b} \\ &= \sigma_X^2 + \sigma_{\hat{X}}^2 - 2\sigma_X^2 \{0 \quad e^{-1/4}\} \begin{Bmatrix} e^{-2/4} \\ e^{-1/4} \end{Bmatrix} \\ &= (500)^2 \left(1 + e^{-2/4} - 2e^{-2/4}\right) \\ &= (500)^2 \left(1 - e^{-2/4}\right) \end{aligned} \tag{34}$$

The standard deviation of the estimator error is thus $\sigma_{_E} = 500\sqrt{1-e^{-2/4}} = 313.6$ *ppm. Note that this is less than the variability of the estimator itself and significantly less than the variability of* X*, due to the restraining effect of correlation between points.*

To compute the required probability, we need to assume a distribution for the random variable $(X_3 - \hat{X}_3)$*. Let us suppose that* X *is normally distributed. Since the estimate* $\hat{X}$ *is simply a sum of* X*'s, it too must be normally distributed, which in turn implies that the quantity* $(X_3 - \hat{X}_3)$ *is normally distributed. We need only specify its mean and standard deviation, then, to fully describe its distribution.*

We saw above that since $\hat{X}_3$ *is an unbiased estimate of* X_3 *that*

$$\mathrm{E}\Big[X_3 - \hat{X}_3\Big] = 0 \tag{35}$$

so that $\mu_{_E} = 0$*. We have just computed the standard deviation of* $(X_3 - \hat{X}_3)$ *as* $\sigma_{_E} = 313.6$ *ppm. Thus,*

$$\mathrm{P}\Big[X_3 - \hat{X}_3 > 400\Big] = \mathrm{P}\left[Z > \frac{400-0}{313.6}\right] = 1 - \Phi(1.28) = 0.1003 \tag{36}$$

3 Geostatistics: Kriging

Kriging is basically best linear unbiased estimation with the added ability to estimate certain aspects of the mean trend. We will re-introduce the topic from the point of view of geostatistics (or kriging) in this section, recognizing that some concepts will be repeated. The application will be to a settlement problem in geotechnical engineering.

The purpose of Kriging is to provide a best estimate of a random field between known data. The basic idea is to estimate $X(\underset{\sim}{x})$ at any point using a weighted linear combination of the values of X at each observation point. Suppose that $X_1, X_2, \ldots, X_n$ are observations of the random field, $X(\underset{\sim}{x})$, at the points $\underset{\sim}{x}_1, \underset{\sim}{x}_2, \ldots, \underset{\sim}{x}_n$. Then the kriged estimated of $X(\underset{\sim}{x})$ at $\underset{\sim}{x}$ is given by

$$\hat{X}(\underset{\sim}{x}) = \sum_{i=1}^{n} \beta_i X_i \tag{37}$$

where the n unknown weights β_i are to be determined to find the best estimate at the point $\underset{\sim}{x}$. It seems reasonable that if the point $\underset{\sim}{x}$ is particularly close to one of the observations, say X_k, then the weight, β_k, associated with X_k would be high. However, if $X(\underset{\sim}{x})$ and X_k are in different (independent) soil layers, for example, then perhaps β_k should be small. Rather than using distance to determine the weights in Eq. (37), it is better to use covariance (or correlation) between the two points since this reflects not only distance but also the effects of differing geologic units, etc.

If the mean can be expressed as in a regression analysis,

$$\mu_X(\underset{\sim}{x}) = \sum_{k=1}^{M} a_k g_k(\underset{\sim}{x}) \tag{38}$$

then the unknown weights can be obtained from the matrix equation

$$\underset{\approx}{K}\underset{\sim}{\beta} = \underset{\sim}{M} \tag{39}$$

where $\underset{\approx}{K}$ and $\underset{\sim}{M}$ depend on the covariance structure,

$$\underset{\approx}{K} = \begin{bmatrix} C_{11} & C_{12} & \cdot & \cdot & \cdot & C_{1n} & g_1(\underset{\sim}{x}_1) & g_2(\underset{\sim}{x}_1) & \cdot & \cdot & \cdot & g_M(\underset{\sim}{x}_1) \\ C_{21} & C_{22} & \cdot & \cdot & \cdot & C_{2n} & g_1(\underset{\sim}{x}_2) & g_2(\underset{\sim}{x}_2) & \cdot & \cdot & \cdot & g_M(\underset{\sim}{x}_2) \\ \cdot & \cdot & \cdot & & & \cdot & \cdot & \cdot & \cdot & & & \cdot \\ \cdot & \cdot & & \cdot & & \cdot & \cdot & \cdot & & \cdot & & \cdot \\ \cdot & \cdot & & & \cdot & \cdot & \cdot & \cdot & & & \cdot & \cdot \\ C_{n1} & C_{n2} & \cdot & \cdot & \cdot & C_{nn} & g_1(\underset{\sim}{x}_n) & g_2(\underset{\sim}{x}_n) & \cdot & \cdot & \cdot & g_M(\underset{\sim}{x}_n) \\ g_1(\underset{\sim}{x}_1) & g_1(\underset{\sim}{x}_2) & \cdot & \cdot & \cdot & g_1(\underset{\sim}{x}_n) & 0 & 0 & \cdot & \cdot & \cdot & 0 \\ g_2(\underset{\sim}{x}_1) & g_2(\underset{\sim}{x}_2) & \cdot & \cdot & \cdot & g_2(\underset{\sim}{x}_n) & 0 & 0 & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & & & \cdot & \cdot & \cdot & \cdot & & & \cdot \\ \cdot & \cdot & & \cdot & & \cdot & \cdot & \cdot & & \cdot & & \cdot \\ \cdot & \cdot & & & \cdot & \cdot & \cdot & \cdot & & & \cdot & \cdot \\ g_M(\underset{\sim}{x}_1) & g_M(\underset{\sim}{x}_2) & \cdot & \cdot & \cdot & g_M(\underset{\sim}{x}_n) & 0 & 0 & \cdot & \cdot & \cdot & 0 \end{bmatrix} \tag{40}$$

in which C_{ij} is the covariance between X_i and X_j and

$$\underset{\sim}{\beta} = \begin{Bmatrix} \beta_1 \\ \beta_2 \\ \cdot \\ \cdot \\ \cdot \\ \beta_n \\ -\eta_1 \\ -\eta_2 \\ \cdot \\ \cdot \\ \cdot \\ -\eta_M \end{Bmatrix} \qquad \underset{\sim}{M} = \begin{Bmatrix} C_{1x} \\ C_{2x} \\ \cdot \\ \cdot \\ \cdot \\ C_{nx} \\ g_1(\underset{\sim}{x}) \\ g_2(\underset{\sim}{x}) \\ \cdot \\ \cdot \\ \cdot \\ g_M(\underset{\sim}{x}) \end{Bmatrix} \tag{41}$$

The quantities η_i are a set of Lagrangian parameters used to solve the variance minimization problem subject to non-bias conditions. Beyond allowing for a solution to the above system of equations, they will be ignored in this simple treatment. The

covariance C_{ix} appearing in the RHS vector $\underset{\sim}{M}$ is the covariance between the i^{th} observation point and the point $\underset{\sim}{x}$ at which the best estimate is to be calculated.

Note that the matrix $\underset{\approx}{K}$ is purely a function of the observation point locations and their covariances – thus it can be inverted once and then Eqs. (39) and (37) used repeatedly at different spatial points to build up the field of best estimates (for each spatial point, the RHS vector $\underset{\sim}{M}$ changes, as does the vector of weights, $\underset{\sim}{\beta}$).

The Kriging method depends upon two things; 1) knowledge of how the mean varies functionally with position, i.e. $g_1, g_2, \ldots$ need to be specified, and 2) knowledge of the covariance structure of the field. Usually, assuming a mean which is either constant ($M = 1$, $g_1(\underset{\sim}{x}) = 1$, $a_1 = \mu_x$) or linearly varying is sufficient. The correct order can be determined by

1. plotting the results and visually checking the mean trend, or by
2. performing a regression analysis, or by
3. performing a more complex structural analysis – see *Mining Geostatistics* by Journel and Huijbregts (Academic Press, 1978) for details on this approach.

The covariance structure can be estimated by the methods discussed in the next chapter, if sufficient data is available, and used directly in Eq. (39) to define $\underset{\approx}{K}$ and $\underset{\sim}{M}$ (with, perhaps some interpolation for covariances not directly estimated). In the absence of sufficient data, a simple functional form for the covariance function is often assumed. A typical model is the Markovian one in which the covariance decays exponentially with separation distance $\tau_{ij} = |\underset{\sim}{x}_1 - \underset{\sim}{x}_2|$;

$$C_{ij} = \sigma_x^2 \exp\left\{-\frac{2|\tau_{ij}|}{\theta}\right\} \tag{42}$$

As mentioned previously, the parameter θ is called the *scale of fluctuation*. Such a model now requires only the estimation of two parameters, σ_x and θ, but assumes that the field is *isotropic* and *statistically homogeneous*. Non-isotropic models are readily available and often appropriate for soils which display layering.

3.1 Estimator Error

Associated with any estimate of a random process derived from a finite number of observations is an estimator error. This error can be used to assess the accuracy of the estimate. Defining the error as the difference between the estimate, $\hat{X}(\underset{\sim}{x})$, and its true (but unknown and random) value, $X(\underset{\sim}{x})$, the estimator mean and corresponding error variance are given by

$$\begin{aligned}
\mu_{\hat{X}}(\underset{\sim}{x}) &= \mathrm{E}\Big[\hat{X}(\underset{\sim}{x})\Big] = \mathrm{E}\big[X(\underset{\sim}{x})\big] = \mu_X(\underset{\sim}{x}) \\
\sigma_E^2 &= \mathrm{E}\left[\Big(\hat{X}(\underset{\sim}{x}) - X(\underset{\sim}{x})\Big)^2\right] = \sigma_X^2 + \underset{\sim}{\beta}_n^T(\underset{\approx}{K}_{n\times n}\underset{\sim}{\beta}_n - 2\underset{\sim}{M}_n)
\end{aligned} \tag{43}$$

where $\underset{\sim}{\beta}_n$ and $\underset{\sim}{M}_n$ are the first n elements of $\underset{\sim}{\beta}$ and $\underset{\sim}{M}$ defined in the previous section, and $\underset{\approx}{K}_{n\times n}$ is the $n \times n$ upper left submatrix of $\underset{\approx}{K}$ containing the covariances, also defined in the previous section. Note that $\hat{X}(\underset{\sim}{x})$ can also be viewed as the conditional mean of $X(\underset{\sim}{x})$ at the point $\underset{\sim}{x}$. The conditional variance at the point $\underset{\sim}{x}$ would then be σ^2_{E}.

3.2 Example: Foundation Consolidation Settlement

Consider the estimation of consolidation settlement under a footing at a certain location given that soil samples/tests have been obtained at 4 neighboring locations. Fig. 1 shows a plan view of the footing and sample locations.

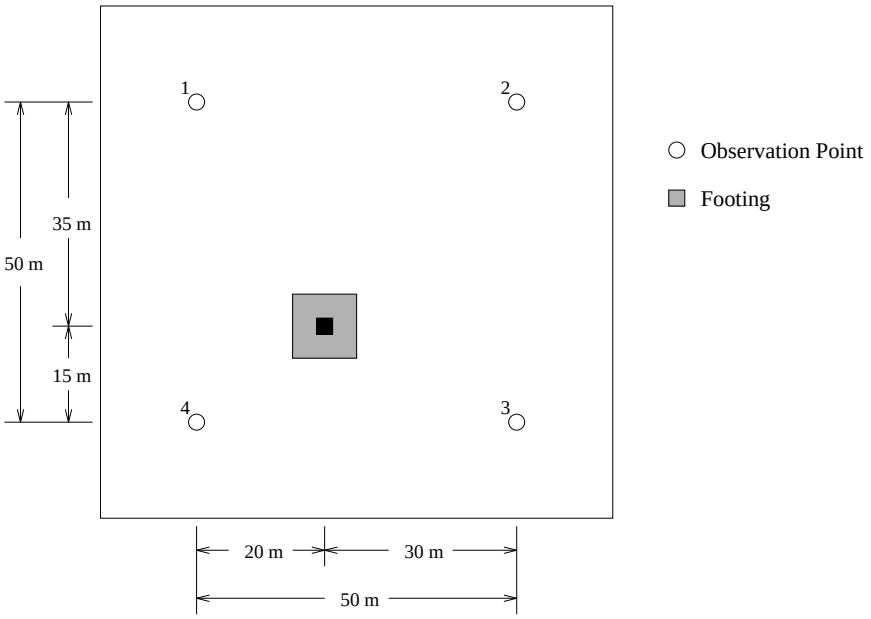


Figure 1. Consolidation settlement plan view with sample points.

The samples and local stratigraphy are used to estimate the soil parameters C_c, e_o, H, and p_o appearing in the consolidation settlement equation

$$S = N\left(\frac{C_c}{1+e_o}\right) H \log_{10}\left(\frac{p_o+\Delta p}{p_o}\right) \tag{44}$$

at each of the sample locations. Each of these 4 parameters are then treated as spatially varying and random between observation points. It is assumed that the estimation error in obtaining the parameters from the samples is negligible compared to field variability, and so this source of uncertainty will be ignored. The model error parameter, N, is assumed an ordinary random variable (not a random field) with mean 1.0 and standard deviation 0.1. The increase in pressure at mid-depth of the clay layer, Δp depends on

the load applied to the footing. We will assume that $\mathrm{E}\big[\Delta p\big] = 0.5$ ksf with standard deviation 0.1.

The task now is to estimate the mean and standard deviation of C_c, e_o, H, and p_o at the footing location using the neighboring observations. Table 1 lists the soil settlement properties obtained at each of the 4 sample points.

In Table 1, we have assumed that all 4 random fields are stationary, with spatially constant mean and variance, the limited data not clearly indicating otherwise. In order to obtain a Kriging estimate at the footing location, we need to establish a covariance structure for the field. Obviously 4 sample points is far too few to yield even a rough approximation of the covariance between samples, especially in two dimensions. Let us assume that experience with similar sites and similar materials leads us to estimate a scale of fluctuation of about 60 m using an exponentially decaying correlation function. That is, we assume that the correlation structure is reasonably well approximated by

$$\rho(\underset{\sim}{x}_i, \underset{\sim}{x}_j) = \exp\left\{-\frac{2}{60}|\underset{\sim}{x}_i - \underset{\sim}{x}_j|\right\} \tag{45}$$

Table 1 Derived soil sample settlement properties.

Sample Point	C_c	e_o	H (inches)	p_o (ksf)
1	0.473	1.42	165	3.90
2	0.328	1.08	159	3.78
3	0.489	1.02	179	3.46
4	0.295	1.24	169	3.74
μ	0.396	1.19	168	3.72
σ^2	0.009801	0.03204	70.56	0.03460

In so doing, we are assuming that the clay layer is horizontally isotropic, also a reasonable assumption. This yields the following correlation matrix between sample points;

$$\underset{\approx}{\rho} = \begin{bmatrix} 1.000 & 0.189 & 0.095 & 0.189 \\ 0.189 & 1.000 & 0.189 & 0.095 \\ 0.095 & 0.189 & 1.000 & 0.189 \\ 0.189 & 0.095 & 0.189 & 1.000 \end{bmatrix} \tag{46}$$

Furthermore, it is reasonable to assume that the same scale of fluctuation applies to all 4 soil properties. Thus, the covariance matrix associated with the property C_c between

sample points is just $\sigma^2_{C_c}\underset{\approx}{\rho} = 0.009801\underset{\approx}{\rho}$. Similarly, the covariance matrix associated with e_o is its variance ($\sigma^2_{e_o} = 0.03204$) times the correlation matrix, etc.

In the following, we will obtain Kriging estimates from each of the 4 random fields ($C_c(\underset{\sim}{x})$, $e_o(\underset{\sim}{x})$ independently. Note that this does not imply that the estimates will be independent, since if the sample properties are themselves correlated, which they most likely are, then the estimates will also be correlated. It is believed that this is a reasonably good approximation given the level of available data. If more complicated cross-correlation structures are known to exist, and have been estimated, the method of *co-Kriging* can be applied – this essentially amounts to the use of a much larger covariance (Kriging) matrix and the consideration of all four fields simultaneously. Co-Kriging also has the advantage of also ensuring that the error variance is properly minimized. However, co-Kriging is not implemented here, since the separate Kriging preserves reasonably well any existing point-wise cross-correlation between the fields and since little is generally known about the actual cross-correlation structure.

The Kriging matrix associated with the clay layer thickness H is then

$$\underset{\approx}{K}_H = \begin{bmatrix} 70.56 & 13.33 & 6.682 & 13.33 & 1 \\ 13.33 & 70.56 & 13.33 & 6.682 & 1 \\ 6.682 & 13.33 & 70.56 & 13.33 & 1 \\ 13.33 & 6.682 & 13.33 & 70.56 & 1 \\ 1 & 1 & 1 & 1 & 0 \end{bmatrix} \tag{47}$$

where, since we assumed stationarity, $M = 1$ and $g_1(\underset{\sim}{x}) = 1$ in Eq. (38). Placing the coordinate axis origin at sample location 4 gives the footing coordinates $\underset{\sim}{x} = (20, 15)$. Thus, the right hand side vector $\underset{\sim}{M}$ is

$$\underset{\sim}{M}_H = \begin{Bmatrix} \sigma^2_H \rho(\underset{\sim}{x}_1, \underset{\sim}{x}) \\ \sigma^2_H \rho(\underset{\sim}{x}_2, \underset{\sim}{x}) \\ \sigma^2_H \rho(\underset{\sim}{x}_3, \underset{\sim}{x}) \\ \sigma^2_H \rho(\underset{\sim}{x}_4, \underset{\sim}{x}) \\ 1 \end{Bmatrix} = \begin{Bmatrix} (70.56)(0.2609) \\ (70.56)(0.2151) \\ (70.56)(0.3269) \\ (70.56)(0.4346) \\ 1 \end{Bmatrix} = \begin{Bmatrix} 18.41 \\ 15.18 \\ 23.07 \\ 30.67 \\ 1 \end{Bmatrix} \tag{48}$$

Solving the matrix equation $\underset{\approx}{K}_H \underset{\sim}{\beta}_H = \underset{\sim}{M}_H$ gives the following four weights (ignoring the Lagrange parameter);

$$\underset{\sim}{\beta}_H = \begin{Bmatrix} 0.192 \\ 0.150 \\ 0.265 \\ 0.393 \end{Bmatrix} \tag{49}$$

in which we can see that the samples which are closest to the footing are most heavily weighted (more specifically, the samples which are most highly correlated with the footing location are the most heavily weighted), as would be expected.

Since the underlying correlation matrix is identical for all 4 soil properties, the weights will be identical for all 4 properties, thus the best estimates at the footing are

$$\begin{aligned}
\hat{C}_c &= (0.192)(0.473) + (0.150)(0.328) + (0.265)(0.489) + (0.393)(0.295) = 0.386 \\
\hat{e}_o &= (0.192)(1.42) + (0.150)(1.08) + (0.265)(1.02) + (0.393)(1.24) = 1.19 \\
\hat{H} &= (0.192)(165) + (0.150)(159) + (0.265)(179) + (0.393)(169) = 169 \\
\hat{p}_o &= (0.192)(3.90) + (0.150)(3.78) + (0.265)(3.46) + (0.393)(3.74) = 3.70
\end{aligned}$$

The estimation errors are given by the equation

$$\sigma_{_E}^2 = \sigma_{_X}^2 + \underset{\sim}{\beta}_n^T (\underset{\approx}{\underset{\sim}{K}}_{n\times n} \underset{\sim}{\beta}_n - 2\underset{\sim}{M}_n) \tag{50}$$

Since the $n \times n$ submatrix of $\underset{\approx}{\underset{\sim}{K}}$ is just the correlation matrix times the appropriate variance, and similarly $\underset{\sim}{M}_n$ is the correlation vector (between samples and footing) times the appropriate variance, the error can be rewritten

$$\sigma_{_E}^2 = \sigma_{_X}^2 \left(1 + \underset{\sim}{\beta}_n^T (\underset{\approx}{\underset{\sim}{\rho}} \underset{\sim}{\beta}_n - 2\underset{\sim}{\rho}_x)\right) \tag{51}$$

where $\underset{\sim}{\rho}_x$ is the vector of correlation coefficients between the samples and the footing (see the calculation of $\underset{\sim}{M}_{_H}$ above). For the Kriging weights and given correlation structure, this yields

$$\sigma_{_E}^2 = \sigma_{_X}^2 (0.719) \tag{52}$$

which gives the following individual estimation errors;

$$\begin{aligned}
\sigma_{_{C_c}}^2 &= (0.009801)(0.719) = 0.00705 \quad \rightarrow \quad \sigma_{_{C_c}} = 0.0839 \\
\sigma_{_{e_o}}^2 &= (0.03204)(0.719) = 0.0230 \quad \rightarrow \quad \sigma_{_{e_o}} = 0.152 \\
\sigma_{_H}^2 &= (70.56)(0.719) = 50.7 \quad \rightarrow \quad \sigma_{_H} = 7.12 \\
\sigma_{_{p_o}}^2 &= (0.03460)(0.719) = 0.0249 \quad \rightarrow \quad \sigma_{_{p_o}} = 0.158
\end{aligned}$$

In summary, then, the variables entering the consolidation settlement formula have the following statistics based on the preceding Kriged estimates;

Variable	Mean	SD	CV
N	1.0	0.1	0.1
C_c	0.386	0.0839	0.217
e_o	1.19	0.152	0.128
H	169	7.12	0.042
p_o	3.70 ksf	0.158	0.043
Δp	0.50 ksf	0.100	0.20

where CV stands for the coefficient of variation.

A first-order approximation to the settlement, via Eq. (44), is thus

$$\mu_S = (1.0)\left(\frac{0.386}{1+1.19}\right)(169)\log_{10}\left(\frac{3.7+0.5}{3.7}\right) = 1.64 \tag{53}$$

To estimate the settlement coefficient of variation, a first order approximation yields,

$$CV_S^2 = \sum_{j=1}^{m}\left(\frac{\partial S}{\partial X_j}\frac{\mu_{X_j}}{\mu_S}\right)_\mu^2 CV_j^2 = \sum_{j=1}^{m} S_j^2 CV_j^2 \tag{54}$$

where the subscript μ on the derivative implies that it is evaluated at the mean of all random variables and CV_j is the coefficient of variation of the j^{th} variable – the variable X_j is replaced by each of N, C_c, etc., in turn. Evaluation of the derivatives at the mean leads to the following table;

X_j	μ_{X_j}	CV_j	S_j	$S_j^2\,CV_j^2$
N	1.0	0.100	1.0	0.01
C_c	0.386	0.217	1.0	0.0471
e_o	1.19	0.128	-0.54	0.0048
H	169	0.042	1.0	0.0018
p_o	3.70	0.043	-0.94	0.0016
Δp	0.50	0.200	0.94	0.0353

so that

$$CV_S^2 = \sum_{j=1}^{m} S_j^2 CV_j^2 = 0.10057 \tag{55}$$

giving a coefficient of variation for the settlement at the footing of 0.317. This is roughly a 10% decrease from the result obtain without the benefit of any neighboring observations. Although this does not seem significant in light of the increased complexity of the above calculations, it needs to be remembered that the contribution to overall uncertainty coming from N and Δp amounts to over 40%. Thus, the coefficient of variation, CV_S, will decrease towards it's minimum (barring improved information about N and/or Δp) of 0.212 as more observations are used and/or observations are taken closer to the footing. For example, if a fifth sample were taken midway between the other 4 samples (at the center of Fig. 8.1), then the variance of each estimator decreases by a factor of 0.46 from the point variance (rather than the factor of 0.719 found above) and the settlement CV becomes 0.285. Note that the reduction in variance can be found prior to actually performing the sampling since the estimator variance depends only on the covariance structure and the assumed functional form for the mean. Thus, the Kriging technique can also be used to plan an optimal sampling scheme – sample points are selected so as to minimize the estimator error.

Once the random field model has been defined for a site, there are ways of analytically obtaining probabilities associated with design criteria, such as the probability of failure. For example, by assuming a normal or lognormal distribution for the footing settlement in the previous section, one can easily estimate the probability that the footing will exceed a certain settlement given it's mean and standard deviation. Assuming the footing settlement to be normally distributed with mean 1.64 inches and a CV of 0.317 (standard deviation $= (0.317)(1.64) = 0.52$) then the probability that the settlement will exceed 2.5 inches is

$$\mathrm{P}\big[S > 2.5\big] = 1 - \Phi\left(\frac{2.5 - 1.64}{0.52}\right) = 1 - \Phi(1.65) = 0.05 \qquad (56)$$

Simulation

Gordon A. Fenton

Dalhousie University, Canada

1 Introduction

Stochastic problems are often very complicated, requiring overly simplistic assumptions in order to obtain closed-form (or exact) solutions. This is particularly true of many geotechnical problems where we don't even have exact analytical solutions to the deterministic problem. For example, general seepage problems, settlement under rigid footings, bearing capacity, and pile capacity problems all lack exact analytical solutions and discussion is ongoing about the various approximations which have been developed over the years. Needless to say, when spatial randomness is added to the problem, even the approximate solutions are often unwieldy, if they can be found at all. For example, one of the simpler problems in geotechnical engineering is that of 'D'Arcy Law' seepage through a clay barrier. If the barrier has a large area, relative to its thickness, and flow is through the thickness, then a 1-D seepage model is appropriate. In this case, a closed-form analytical solution to the seepage problem is available. However, if the clay barrier has spatially variable hydraulic conductivity, then the 1-D model is no longer appropriate (flow lines avoid low conductivity regions) and even the deterministic problem no longer has a simple closed form solution. Problems of this type, and most other geotechnical problems, are best tackled through simulation. *Simulation* is the process of producing reasonable replications of the real world in order to study the probabilistic nature of the response to the real world. In particular, simulations allow the investigation of more realistic geotechnical problems, potentially yielding entire probability distributions related to the output quantities of interest. A simulation basically proceeds by the following steps;

1. by taking as many observations from the 'real world' as are feasible, the stochastic nature of the 'real world' problem can be estimated. From the raw data, histogram(s), statistical estimators, and goodness-of-fit tests, a distribution with which to model the problem is decided upon. Pertinent parameters, such as the mean, variance, scale of fluctuation, occurrence rate, etc., may be of interest in characterizing the randomness (see Chapter 5),

2. a random variable or field, following the distribution decided upon in the previous step, is defined,
3. a *realization* of the random variable/field is generated using a *pseudo-random number generator* or a *random field generator*,
4. the response of the system to the random input generated in the previous step is evaluated,
5. the above algorithm is repeated from step (3) for as many times as are feasible, recording the responses and/or counting the number of occurrences of a particular response observed along the way.

This process is called *Monte Carlo simulation*, after the famed randomness of the gambling houses of Monte Carlo. The probability of any particular system response can now be estimated by dividing the number of occurrences of that particular system response by the total number of simulations. In fact, if all of the responses are retained in numerical form, then a histogram of the responses forms an estimate of the probability distribution of the system response. Thus, Monte Carlo simulations are a powerful means of obtaining probability distribution estimates for very complex problems. Only the response of the system to a known, *deterministic*, input needs to be computed at each step during the simulation. In addition, the above methodology is easily extended to multiple independent random variables or fields – in this case the distribution of each random variable or field needs to be determined in step (1) and a realization for each generated in step (3). If the multiple random variables or fields are not independent, then the process is slightly more complicated and will be considered in the context of random fields in the second part of this chapter.

Monte Carlo simulations essentially replicate the experimental process, and are representative of the experimental results. The accuracy of the representation depends entirely on how accurately the fitted distribution matches the experimental process (e.g., how well the distribution matches the random field of soil properties). The outcomes of the simulations can be treated statistically, just as any set of observations can be treated. As with any statistic, the accuracy of the method generally increases as the number of simulations increases.

In theory, simulation methods can be applied to large and complex systems and often the rigid idealizations and/or simplifications necessary for analytical solutions can be removed, resulting in more realistic models. However, in practice, Monte Carlo simulations may be limited by constraints of economy and computer capability. Moreover, solutions obtained from simulations may not be amenable to generalization or extrapolation. Therefore, as a general rule, Monte Carlo methods should be used only as a last resort: that is, when and if analytical solution methods are not available or are ineffective (eg. because of gross idealizations). Monte Carlo solutions are also often a means of verifying or validating approximate analytical solution methods.

One of the main tasks in Monte Carlo simulation is the generation of random numbers having a prescribed probability distribution. Uniformly distributed random number

generation will be studied in Section 2. Some techniques for generating random variates from other distributions will be seen in Section 3. Finally, techniques of generating random fields are considered starting in Section 4.

2 Random Number Generators

2.1 Common Generators

Recall that the $U(0,1)$ distribution is a continuous uniform distribution on the interval from zero to one. Any one number in the range is just as likely to turn up as any other number in the range. For this reason, the continuous uniform distribution is the simplest of all continuous distributions. While techniques exist to generate random variates from other distributions, they all employ $U(0,1)$ random variates. Thus, if a good uniform random number generator can be devised, its output can also be used to generate random numbers from other distributions (eg. exponential, Poisson, normal, ...), which can be accomplished by an appropriate *transformation* of the uniformly distributed random numbers.

Most of the best and most commonly used uniform random number generators are so-called *arithmetic generators*. These employ sequential methods where each number is determined by one or several of its predecessors according to a fixed mathematical formula. If carefully designed, such generators can produce numbers that *appear* to be independent random variates from the $U(0,1)$ distribution, in that they pass a series of statistical tests (to be discussed shortly). In the sense that sequential numbers are not truly random, being derived from previous numbers in some deterministic fashion, these generators are often called *pseudorandom number generators*.

A "good" arithmetic uniform random number generator should possess several properties:

1. the numbers generated should appear to be independent and uniformly distributed,
2. the code should be fast and not require large amounts of storage
3. have the ability to reproduce a given stream of random numbers exactly
4. should have a very long period.

The ability to reproduce a given stream of random numbers is sometimes useful when attempting to compare the responses of two different systems (or designs) to random input. If the input is not the same to the two systems, then their responses will be naturally different, and it is more difficult to determine how the systems actually differ. Being able to 'feed' the two systems the same stream of random numbers allows the system differences to be directly studied.

The most popular arithmetic generators are *linear congruential generators* (LCGs) first introduced by Lehmer (1951). In this method, a sequence of integers $Z_1, Z_2, ...$ are defined by the recursive formula

$$Z_i = (aZ_{i-1} + c)(\text{mod } m) \tag{1}$$

where m is the modulus, a is a multiplier, c is an increment, and all three parameters are positive integers. The sequence starts by computing Z_1 using Z_0, where Z_0 is a positive integer *seed* or starting value. Effectively, Eq. 1 sets Z_i to the *remainder* when $aZ_{i-1} + c$ is divided by m. Since the resulting Z_i must lie from 0 to $m - 1$, we can obtain a $[0, 1)$ uniformly distributed U_i by setting $U_i = Z_i/m$ – the $[0, 1)$ notation means that U_i can be 0, but cannot be 1. The largest value that U_i can take is $(m - 1)/m$, which can be quite close to 1 if m is large. Also, because Z_i can only take on m different possible values, U_i can only take on m possible values between 0 and 1. Namely, U_i can have values $0, 1/m, 2/m, \ldots, (m - 1)/m$. In order for U_i to appear continously uniformly distributed on $[0, 1)$, then, m should be selected to be a large number. In addition a, c, and Z_o should all be less than m.

One sees immediately from Eq. 1 that the sequence of Z_i are completely dependent; Z_1 is obtained from Z_0, Z_2 is obtained from Z_1, and so on. For fixed values of a, c, and m, the same sequence of Z_i values will always be produced for the same starting seed, Z_0. Thus, Eq. 1 can reproduce a given stream of pseudorandom numbers exactly, so long as the starting seed is known. But will the derived U_i appear independent and uniformly distributed? It turns out that if a, c, and m are correctly selected, the sequence of U_i will appear to be largely independent and uniformly distributed.

One may also notice that if $Z_0 = 3$ produces $Z_1 = 746$, then whenever $Z_{i-1} = 3$, the next generated value will be $Z_i = 746$. This property results in a very undesirable phenomenon called *periodicity* that quite a number of rather common random number generators suffer from. Suppose that you were unlucky enough to pick a starting seed, say $Z_0 = 83$ on one of these poor random number generators that just happened to yield remainder 83 when $83a + c$ is divided by m. Then $Z_1 = 83$. In fact, the resulting sequence of 'random' numbers will be $\{83, 83, 83, 83, \ldots\}$. We say that this particular stream of random variates has periodicity equal to one.

Why is periodicity to be avoided? To answer this question, let us suppose you are estimating the average of a system's response by simulation. The simulated random input $U_1, U_2, \ldots, U_n$ results in responses $X_1, X_2, \ldots, X_n$. You may then compute the average response as

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{2}$$

and statistical theory tell you that the standard error on this estimate ($\pm$ one standard deviation) is

$$s_{\bar{X}} = \frac{s}{\sqrt{n}} \tag{3}$$

where

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{4}$$

The standard error (Eq. 3) reduces towards zero as n increases, so long as the X_i's are independent. Now, suppose that we set $n = 1,000,000$ and pick a starting seed $Z_0 = 261$. Suppose further that this particular seed results in Z_i, $i = 1, 2, \ldots 10^6$, being the sequence $\{94, 4832, 325, 94, 4832, 325, \ldots\}$ with periodicity 3. Then, instead of $1,000,000$ independent input values, as assumed, we actually only have 3 independent values, each repeated 333,333 times. Not only have we wasted a lot of computer time, but our estimate of the average system response might be very much in error – we assume that its standard error is $s/\sqrt{10^6} = 0.001s$, whereas it is actually $s/\sqrt{3} = 0.6s$ – 600 times less accurate than we had thought!

Example 1:

What are the first three random numbers produced by the linear congruental generator

$$Z_i = (25Z_{i-1} + 55)(\text{mod } 96) \tag{5}$$

for starting seed $Z_0 = 21$.

Solution:

Since the modulus is 96, the interval $[0, 1)$ *will be subdivided only into at most 96 possible 'random' values. Normally, the modulus is taken to be much larger to give a fairly fine resolution on the unit interval. However, with* $Z_0 = 21$ *we get*

$$\begin{aligned}
Z_1 &= (25(21) + 55)(\text{mod } 96) \\
&= 580(\text{mod}96) \\
&= 4 \\
Z_2 &= (25(4) + 55)(\text{mod } 96) \\
&= 155(\text{mod}96) \\
&= 59 \\
Z_3 &= (25(59) + 55)(\text{mod } 96) \\
&= 1530(\text{mod}96) \\
&= 90
\end{aligned}$$

so that $U_1 = 4/96 = 0.042$*,* $U_2 = 59/96 = 0.615$*, and* $U_3 = 90/96 = 0.938$.

The maximum periodicity an LCG such as Eq. 1 can have is m, and this will occur only if a, c, and m are selected very carefully. We say that a generator has *full period* if its period is m. A generator which is full period will produce exactly one of each possible value, $\{0, 1, \ldots, m-1\}$, in each cycle. If the generator is good, all of these possible values will appear to occur in random order.

To help us choose the values of m, a, and c so that the generator has full period, the following theorem, proved by Hull and Dobell (1962) is valuable

Theorem 1. The LCG defined by Eq. 1 has full period if and only if the following three conditions hold:

a) The only positive integer that exactly divides both m and c is 1.

b) If q is a prime number (divisible only by itself and 1) that divides m, then q divides $a - 1$.

c) If 4 divides m, then 4 divides $a - 1$.

Condition (b) must be true of all prime factors of m. For example, $m = 96$ has two prime factors, 2 and 3, not counting 1. If $a = 25$, then $a - 1 = 24$ is divisible by both 2 and 3, so that condition (b) is satisfied. In fact, it is easily shown that the LCG $Z_i = (25Z_{i-1} + 55)(\mathrm{mod}\ 96)$ used in the previous example is a full period generator.

Park and Miller (1988) proposed a "Minimal Standard" (MS) generator with constants

$$a = 7^5 = 16807, \qquad c = 0, \qquad m = 2^{31} - 1 = 2147483647$$

which has a periodicity of $m - 1$ or about 2×10^9. The only requirement is that the seed 0 must never be used. This form of the LCG, that is having $c = 0$, is called a *multiplicative* LCG,

$$Z_{i+1} = aZ_i(\mathrm{mod}\ m) \tag{6}$$

which has a small efficiency advantage over the general LCG of Eq. 1 since the addition of c is no longer needed. However, most modern CPU's are able to do a vector multiply and add simultaneously, so this efficiency advantage is probably non-existent. Multiplicative LCGs can no longer be full period because m now exactly divides both m and $c = 0$. However, a careful choice of a and m can lead to a period of $m - 1$ and only zero is excluded from the set of possible Z_i values – in fact, if zero is not excluded from the set of possible results of Eq. 6, then the generator will eventually just return zeroes. That is, once $Z_i = 0$ in Eq. 6, it remains zero forever. The constants selected by Park and Miller (1988) for the MS generator achieves a period of $m - 1$ and excludes zero. Possible values for U_i using the MS generator are $\{1/m, 2/m, \ldots, (m-1)/m\}$ and so both of the endpoints, 0 and 1, are excluded. Excluding the endpoints is useful for the generation of random variates from other distributions which involves taking the logarithm of U or $(1 - U)$ (since $ln(0) = -\infty$).

When implementing the MS generator on computers using 32 bit integers, the product aZ_i will generally result in an integer overflow. In their *ran0* function, Press et al. (1997) provide a 32 bit integer implementation of the MS generator using a technique developed by Schrage (1979).

One of the main drawbacks to the "Minimal Standard" generator is that there is some correlation between successive values. For example, when Z_i is very small, the product aZ_i can still be very small (relative to m). Thus, very small values are always followed by small values. For example, if $Z_i = 1$, then $Z_{i+1} = 16807$,

$Z_{i+2} = 282475249$. The corresponding sequence of U_i is 4.7×10^{-10}, 7.8×10^{-6}, and 0.132. Any time that U_i is less than 1×10^{-6}, the next value will be less than 0.0168.

To remove the serial correlation in the "Minimal Standard" generator along with this problem of small values following small values, a technique suggested by Bays and Durham and reported by Knuth (1981) is to use two LCG's; one a "Minimal Standard" generator, and the second to randomly shuffle the output from the first. In this way, U_{i+1} is not returned by the algorithm immediately after U_i, but rather at some random time in the future. This effectively removes the problem of serial correlation. In their 2nd Edition of *Numerical Recipes*, Press et al (1997) present a further improvement, due to L'Ecuyer (1988), which involves combining two different pseudorandom sequences, with different periods, as well as applying the random shuffle. The resulting sequence has a period which is the least common multiple of the two periods, which in Press et al's implementation is about 2.3×10^{18}. See Press et al's *RAN2* function, which is what the authors of this book use as their basic random number generator.

3 Generating Non-Uniform Random Variables

The basic ingredient needed for all common methods of generating random variates or random processes (which are sequences of random variables) from any distribution is a sequence of $U(0, 1)$ random variates. It is thus important that the basic random number generator be good. This issue was covered in the previous section, and standard "good" generators are readily available.

For most common distributions, efficient and exact generation algorithms exist that have been thoroughly tested and used over the years. Less common distributions may have several alternative algorithms available. For these, there are a number of issues that should be considered before choosing the best algorithm;

1. *exactness:* unless there is a significant sacrifice in execution time, methods which reproduce the desired distribution exactly, in the limit as $n \to \infty$, are preferable. When only approximate algorithms are available, those which are accurate over the largest range of parameter values are preferable.
2. *execution time:* with modern computers, setup time, storage, and time to generate each variate are not generally a great concern. However, if the number of realizations is to be very large, execution time may be a factor which should be considered.
3. *simplicity:* algorithms which are difficult to understand and implement generally involve significant debug time and should be avoided. All other factors being similar, the simplest algorithm is preferable.

Here, the most important general approaches for the generation of random variates from arbitrary distributions will be examined. A few examples will be presented and

the relative merits of the various approaches will be discussed.

3.1 Methods of Generation

The most common methods used to generate random variates are;

1. inverse transform
2. convolution
3. acceptance-rejection

Of these, the inverse transform and convolution methods are exact, while the acceptance-rejection method is approximate. We will only discuss the exact methods here.

3.1.1 Inverse Transform Method

Consider a continuous random variable X that has cumulative distribution function $F_X(x)$ that is strictly increasing. The normal distribution is an example of a distribution where $F_X(x)$ is strictly increasing for all x. This assumption is invoked to ensure that there is only one value of $F_X(x)$ for each value of x, or, stated in a way more appropriate for this method, there is only one value of x for each $F_X(x)$.

The last means that there is only one value of x for each $F(x)$. In this case, the inverse transform method generates a random variate from F by:

1. Generate $u \sim U(0, 1)$
2. Return $x = F^{-1}(u)$

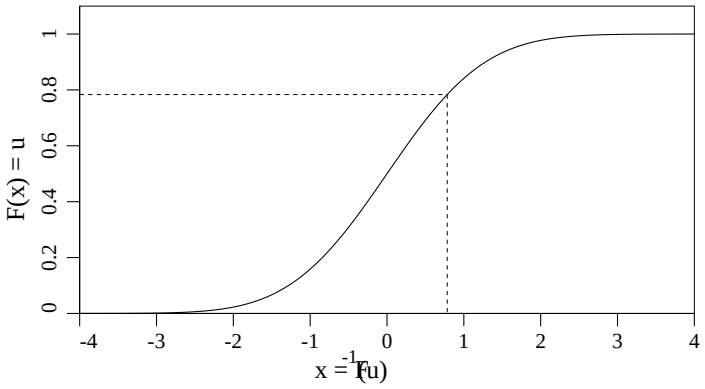


Figure 1. Inverse transform random number generation technique.

Note that $F^{-1}(u)$ will always be defined under the above assumptions since u lies between 0 and 1. Figure 1 illustrates the idea graphically. Since a randomly generated value of U, in this case 0.78, always lies between 0 and 1, the CDF plot can be entered

on the vertical axis, read across to where it intersects $F(x)$, then read down to obtain the appropriate value of x, in this case 0.8. Repetition of this process results in x being returned in proportion to its density, since more 'hits' are obtained where the CDF is the steepest (highest derivative, and, hence, highest density).

The inverse transform method is the best method when the cumulative of the distribution function for generation can be easily "inverted". This includes a number of common distributions, such as the uniform, exponential, Weibull, and Rayleigh.

For example, if X is exponentially distributed, then realizations of X can be obtained by setting

$$F(x) = 1 - e^{-\lambda x} \tag{7}$$

equal to a randomly generated value of U. Setting $u = F(x)$ and inverting gives

$$x = -\frac{\ln(1-u)}{\lambda} \tag{8}$$

Note that since $(1 - U)$ is distributed identically to U, then this can be simplified to

$$x = -\frac{\ln(u)}{\lambda} \tag{9}$$

Admittedly, this leads to a different set of values in the realization, but the *ensemble* of realizations has the same distribution, and that is all that is important. This formulation is also slightly more efficient since one operation has been eliminated. However, which form should be used also depends on the nature of the pseudo-random number generator. Most generators omit either the 0 or the 1, at one of the endpoints of the distribution. Some generators omit both. However, if a generator allows a 0 to occur occasionally, then the form with $\ln(1-u)$ should be used to avoid numerical exceptions ($\ln(0) = -\infty$). Similarly, if a generator allows 1.0 to occur occasionally, then $\ln(u)$ should be used. If both can appear, then the algorithm should specifically guard against an error using if-statements.

The inverse transform approach can also be used on *discrete* random variates, but with a slightly modified algorithm:

1. Generate u from the distribution $\mathrm{U}(0,1)$
2. Determine the smallest x_i such that $F(x_i) \geq u$, and return $x = x_i$.

Another way of stating this algorithm is as follows: since the random variable is discrete, the unit interval can be split up into adjacent subintervals, the first having width equal to $\mathrm{P}\big[X = x_1\big]$, the second having width $\mathrm{P}\big[X = x_2\big]$ and so on. Then assign x according to whichever of these subintervals contains the generated u. There is a computational issue of how to *look* for the subinterval that contains a given u and some approaches are better than others. In particular if x_j, $j = 1, 2, \ldots, m$, are equi-likely outcomes, then $i = \mathrm{int}(1.0 + mu)$, where $\mathrm{int}(\cdot)$ means integer part. This also assumes u can never quite equal 1.0, that is, the generator excludes 1.0 – if 1.0 is possible, then add 0.999999 instead of 1.0 to mu. Now the discrete realization is $x = x_i$.

Both the continuous and discrete versions of the inverse-transform method can be combined, at least formally, to deal with distributions which are *mixed*, ie. having both continuous and discrete components, as well as for continuous distribution functions with flat spots.

Over and above its intuitive appeal, there are three other main advantages to the inverse-transform method:

1. it can easily be modified to generate from truncated distributions,
2. it can be modified to generate order statistics (useful in reliability, or lifetime, applications), and
3. it facilitates variance-reduction techniques (where portions of the CDF are 'polled' more heavily than others, usually in the tails of the distribution, and then resulting statistics corrected to account for the biased 'polling')

The inverse-transform method requires a formula for F^{-1}. However, closed form expressions for the inverse are not known for some distributions, such as the normal, the lognormal, the gamma, and the beta. For such distributions, numerical methods are required to return the inverse. This is the main disadvantage of the inverse-transform method. There are other techniques specifically designed for some of these distributions which will be discussed in the following. In particular, the gamma distribution is often handled by convolution (see next), whereas a simple trigonometric transformation can be used to generate normally distributed variates (and further raising the normal variate to the power e produces a lognormally distributed random variate).

3.1.2 Convolution

The method of *convolution* can be applied when the random variable of interest can be expressed as a sum of other random variables. This is the case for many important distributions – most notably, recall that the Gamma distribution, with integer α, can be expressed as the sum of α exponentially distributed and independent random variables.

For the convolution method, it is assumed that there are independent and identically distributed random variables $Y_1, Y_2, ..., Y_m$ (for fixed m), each with distribution $F(y)$ such that $Y_1 + Y_2 + ... + Y_m$ has the same distribution as X. Hence, X can be expressed as

$$X = Y_1 + Y_2 + ... + Y_m \tag{10}$$

For the method to work efficiently, it is further assumed that random variates for the Y_j's can be generated more readily than X itself directly (otherwise one would not bother with this approach). The convolution algorithm is then quite intuitive:

1. Generate $Y_1, Y_2, ..., Y_m$ i.i.d. each with distribution $F_Y(y)$
2. Return $X = Y_1 + ... + Y_m$.

Note that some other generation method, eg. inverse transform, is required to execute Step 1.

3.2 Generating Common Continuous Random Variates

UNIFORM ON (a, b)

Solving $u = F(x)$ for x yields, for $0 \leq u \leq 1$,

$$x = F^{-1}(u) = a + (b - a)u \tag{11}$$

and the inverse-transform method can be applied as follows;

1. Generate $u \sim U(0, 1)$.
2. Return $x = a + (b - a)u$.

EXPONENTIAL

Solving $u = F(x)$ for x yields, for $0 \leq u \leq 1$,

$$x = F^{-1}(u) = -\frac{\ln(1 - u)}{\lambda} \stackrel{d}{=} -\frac{\ln(u)}{\lambda} \tag{12}$$

where $\stackrel{d}{=}$ implies equivalence in distribution. Now the inverse-transform method can be applied as follows;

1. Generate $u \sim U(0, 1)$.
2. Return $x = -\ln(u)/\lambda$.

GAMMA

Considering the particular form of the Gamma distribution discussed in the Review of Probability Theory Chapter,

$$f_{T_k}(t) = \frac{\lambda\,(\lambda t)^{k-1}}{(k-1)!}\, e^{-\lambda t} \qquad t \geq 0 \tag{13}$$

where T_k is the sum of k independent exponentially distributed random variables, each with mean rate λ. In this case, the generation of random values of T_k proceeds as follows;

1. generate k independent exponentially distributed random variables, $X_1, X_2, \ldots, X_k$, using the algorithm given above,
2. Return $T_k = X_1 + X_2 + \cdots + X_k$

For the more general Gamma distribution, where k is not integer, the interested reader is refered to Law and Kelton (2000).

WEIBULL

Solving $u = F(x)$ for a Weibull distribution yields, for $0 \leq u \leq 1$,

$$x = F^{-1}(u) = \frac{\left[-\ln(1 - u)\right]^{1/\beta}}{\lambda} \stackrel{d}{=} (-\ln u)^{1/\beta}/\lambda \tag{14}$$

and the inverse-transform method can be applied to give

1. Generate $u \sim U(0,1)$.
2. Return $x = (-\ln u)^{1/\beta} / \lambda$

NORMAL

Since neither the normal distribution function nor its inverse has a simple closed-form expression, one must use a numerical method to apply the inverse-transform method. Some packages use the latter however the *transformation* method suggested by Box and Muller (1958) is exact, simple to use and thus much more popular.

First, note that given $X \sim N(0,1)$, the more general random variable $X' \sim N(\mu, \sigma^2)$ is obtained by setting $X' = \mu + \sigma X$. Thus, attention can be restricted to generating from $N(0,1)$, ie. standard random variates. The method is simply as follows

1. Generate $u_1 \sim U(0,1)$ and $u_2 \sim U(0,1)$.
2. Form $x_1 = \sqrt{-2 \ln u_1} \cos(2\pi u_2)$ and $x_2 = \sqrt{-2 \ln u_1} \sin(2\pi u_2)$
3. Return x_1 on this call to the algorithm and x_2 on the next call (so that the whole algorithm is run only on every second call).

The above method generates realizations for X_1 and X_2 which are *independent* $N(0,1)$ random variates.

LOGNORMAL

Recall that the lognormal distribution results from the following: If Y is normally distributed with mean μ and variance σ^2, then e^Y is lognormally distributed with parameters μ and σ^2. The generation algorithm is simple;

1. Generate normally distributed Y with mean μ and variance σ^2 (see previous algorithm).
2. Return $X = e^Y$.

EMPIRICAL

Sometimes a theoretical distribution that fits the data cannot be found. In this case, the observed data may be used directly to specify (in some sense) a usable distribution called an empirical distribution.

For continuous random variables, the type of empirical distribution that can be defined depends on whether the actual values of the individual original observations $x_1, x_2, \ldots, x_n$ are available or only the *number* of x_i's that fall into each of several specified intervals. We will consider the case where all of the original data are available.

Using all of the available observations, a continuous, piecewise-linear distribution function F can be defined by first sorting the x_i's from smallest to largest. Let $x_{(i)}$ denote the i^{th} smallest of the x_j's, so that $x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}$. Then F is defined

by

$$F(x) = \begin{cases} 0 & \text{if } x < x_{(1)} \\ \frac{i-1}{n-1} + \frac{x - x_{(i)}}{(n-1)(x_{(i+1)} - x_{(i)})} & \text{if } x_{(i)} \leq x < x_{(i+1)} \text{ for } i = 1, 2, ..., n-1 \\ 1 & \text{if } x_{(n)} \leq x \end{cases} \tag{15}$$

Since the function $F(x)$ is a series of steps of height 0, $1/(n-1)$, $2/(n-1)$, ..., $(n-2)/(n-1)$, and 1 allows the the generation to conceptually involve generating $u \sim U(0,1)$, figuring out the index i of the step closest to u, and returning $x_{(i)}$. We will actually interpolate between the the step below u and the step above. The following algorithm results;

1. Generate $u \sim U(0,1)$, let $r = (n-1)u$, and let $i = \text{int}(r) + 1$ where $\text{int}(\cdot)$ means integer part.
2. Return $x = x_{(i)} + (r - i + 1)(x_{(i+1)} - x_{(i)})$.

3.2.1 Generating Discrete Random Variates

The discrete inverse-transform methods may also be applied to generate random variables from the more common discrete probability distributions. The fact that these methods use the inverse-transform is not always evident, however in most cases they do.

BERNOULLI

If the probability of 'succcess' is p, then

1. Generate $u \sim U(0,1)$.
2. If $u \leq p$, return $x = 1$. Otherwise, return $x = 0$.

DISCRETE UNIFORM

1. Generate $u \sim U(0,1)$.
2. Return $x = i + \text{int}\big((j - i + 1)u\big)$, where i and j are the upper and lower discrete bounds and $\text{int}(\cdot)$ means the integer part.

BINOMIAL

To generate a binomial distributed random variate with parameters n and p,

1. Generate $y_1, y_2, ..., y_n$ independent Bernoulli random variates, each with parameter p,
2. Return $x = y_1 + y_2 + ... + y_n$.

GEOMETRIC

1. Generate $u \sim U(0,1)$.

2. Return $x = \text{int}\left(\dfrac{\ln u}{\ln(1-p)}\right)$.

NEGATIVE BINOMIAL

If T_m is the number of trials until the m'th success, and T_m follows a negative binomial distribution with parameter p, then T_m can be written as the sum of m geometric distributed random variables. The generation thus proceeds by convolution;

1. Generate $y_1, y_2, \ldots, y_m$ independent geometric random variates, each with parameter p
2. Return $T_m = y_1 + y_2 + \cdots + y_m$

POISSON

If N_t follows a Poisson distribution with parameter $r = \lambda t$, then N_t is the number of 'arrivals' in time interval of length t, where 'arrivals' arrive with mean rate λ. Since interarrival times are independent and exponentially distributed for a Poisson process, we could proceed by generating a series of k exponentially distributed random variables, each with parameter λ, until their sum just exceeds t. Then the realization of N_t is $k-1$; that is, $k-1$ arrivals occurred within time t, the k'th arrival was after time t.

An equivalent and more efficient algorithm was derived by Law and Kelton (2000) by essentially working in the logarithm space to be

1. Let $a = e^{-r}$, $b = 1$, and $i = 0$, where $r = \lambda t$.
2. Generate $u_{i+1} \sim U(0,1)$ and replace b by bu_{i+1}. If $b < a$, return $N_t = i$.
3. Replace i by $i+1$ and go back to step 2.

4 Generating Random Fields

Random field models of complex engineering systems having spatially variable properties are becoming increasingly common. This trend is motivated by the widespread acceptance of reliability methods in engineering design and is made possible by the increasing power of personal computers. It is no longer sufficient to base designs on best estimate or mean values alone. Information quantifying uncertainty and variability in the system must also be incorporated to allow the calculation of failure probabilities associated with various limit state criteria. To accomplish this, a probabilistic model is required. In that most engineering systems involve loads and materials spread over some spatial extent, their properties are appropriately represented by random fields. For example, to estimate the failure probability of a highway bridge, a designer may represent both concrete strength and input earthquake ground motion using independent random fields, the latter time varying. Subsequent analysis using a Monte Carlo approach and a dynamic finite element package would lead to the desired statistics.

In the remainder of this chapter, a number of different algorithms which can be used to produce scalar multi-dimensional random fields are evaluated in light of their accuracy, efficiency, ease of implementation, and ease of use. Many different random field generator algorithms are available of which the following are perhaps the most common:

1. Moving Average (MA) methods,
2. Covariance Matrix Decomposition,
3. Discrete Fourier Transform (DFT) method,
4. Fast Fourier Transform (FFT) method,
5. Turning Bands Method (TBM),
6. Local Average Subdivision (LAS) method,

In all of these methods, only the first two moments of the target field may be specified, namely the mean and covariance structure. Since this completely characterizes a Gaussian field, attention will be restricted in the following to such fields. Non-Gaussian fields may be created through non-linear transformations of Gaussian fields, however some care must be taken since the mean and covariance structure will also be transformed. In addition, only weakly homogeneous fields, whose first two moments are independent of spatial position, will be considered here.

The FFT, TBM and LAS methods are typically much more efficient than the first three methods discussed above. However, the gains in efficiency do not always come without some loss in accuracy, as is typical in numerical methods. In the next few subsections, implementation strategies for these methods are presented and the types of errors associated with each method and ways to avoid them will be discussed in some detail. Finally the methods will be compared and guidelines as to their use suggested.

4.1 Moving Average Method

The Moving Average (MA) technique of simulating random processes is a well known approach involving the expression of the process as an average of an underlying white noise process. Formally, if $Z(\underset{\sim}{x})$ is the desired zero mean process (a nonzero mean can always be added on later) then

$$Z(\underset{\sim}{x}) = \int_{-\infty}^{\infty} f(\underset{\sim}{\xi})\, dW(\underset{\sim}{x} + \underset{\sim}{\xi}), \tag{16a}$$

or equivalently,

$$Z(\underset{\sim}{x}) = \int_{-\infty}^{\infty} f(\underset{\sim}{\xi} - \underset{\sim}{x})\, dW(\underset{\sim}{\xi}), \tag{16b}$$

in which $dW(\underset{\sim}{\xi})$ is the incremental white noise process at the location $\underset{\sim}{\xi}$ with statistical properties

$$\begin{aligned} \mathrm{E}\big[dW(\underset{\sim}{\xi})\big] &= 0, \\ \mathrm{E}\big[dW(\underset{\sim}{\xi})^2\big] &= d\underset{\sim}{\xi}, \\ \mathrm{E}\big[dW(\underset{\sim}{\xi})dW(\underset{\sim}{\xi}')\big] &= 0, \quad \text{if } \underset{\sim}{\xi} \neq \underset{\sim}{\xi}', \end{aligned} \tag{17}$$

and $f(\underset{\sim}{\xi})$ is a weighting function determined from the desired second order statistics of $Z(\underset{\sim}{x})$

$$\begin{aligned} \mathrm{E}\big[Z(\underset{\sim}{x})\, Z(\underset{\sim}{x}+\underset{\sim}{\tau})\big] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(\underset{\sim}{\xi}-\underset{\sim}{x})\, f(\underset{\sim}{\xi}'-\underset{\sim}{x}-\underset{\sim}{\tau})\, \mathrm{E}\big[dW(\underset{\sim}{\xi})\, dW(\underset{\sim}{\xi}')\big]\,, \\ &= \int_{-\infty}^{\infty} f(\underset{\sim}{\xi}-\underset{\sim}{x})\, f(\underset{\sim}{\xi}-\underset{\sim}{x}-\underset{\sim}{\tau})\, d\underset{\sim}{\xi}. \end{aligned} \tag{18}$$

If $Z(\underset{\sim}{x})$ is homogeneous, then the dependence on $\underset{\sim}{x}$ disappears, and (18) can be written in terms of the covariance function (note by 17 that $\mathrm{E}\big[Z(\underset{\sim}{x})\big] = 0$),

$$B(\underset{\sim}{\tau}) = \int_{-\infty}^{\infty} f(\underset{\sim}{\xi})\, f(\underset{\sim}{\xi}-\underset{\sim}{\tau})\, d\underset{\sim}{\xi}. \tag{19}$$

Defining the Fourier transform pair corresponding to $f(\underset{\sim}{\xi})$ in n-dimensions to be,

$$F(\underset{\sim}{\omega}) = \frac{1}{(2\pi)^n}\int_{-\infty}^{\infty} f(\underset{\sim}{\xi})\mathrm{e}^{-i\underset{\sim}{\omega}\cdot\underset{\sim}{\xi}}\, d\underset{\sim}{\xi}, \tag{20a}$$

$$f(\underset{\sim}{\xi}) = \int_{-\infty}^{\infty} F(\underset{\sim}{\omega})\mathrm{e}^{i\underset{\sim}{\omega}\cdot\underset{\sim}{\xi}}\, d\underset{\sim}{\omega}, \tag{20b}$$

then by the convolution theorem Eq. 19 can be expressed as

$$B(\underset{\sim}{\tau}) = (2\pi)^n \int_{-\infty}^{\infty} F(\underset{\sim}{\omega})\, F(-\underset{\sim}{\omega})\, \mathrm{e}^{-i\underset{\sim}{\omega}\cdot\underset{\sim}{\tau}}\, d\underset{\sim}{\omega}, \tag{21}$$

from which a solution can be obtained from the Fourier transform of $B(\underset{\sim}{\tau})$,

$$F(\underset{\sim}{\omega})\, F(-\underset{\sim}{\omega}) = \frac{1}{(2\pi)^{2n}}\int_{-\infty}^{\infty} B(\underset{\sim}{\tau})\, \mathrm{e}^{-i\underset{\sim}{\omega}\cdot\underset{\sim}{\tau}}\, d\underset{\sim}{\tau}. \tag{22}$$

Note that the symmetry in the left hand side of (22) comes about due to the symmetry $B(\underset{\sim}{\tau}) = B(-\underset{\sim}{\tau})$. It is still necessary to assume something about the relationship between $F(\underset{\sim}{\omega})$ and $F(-\underset{\sim}{\omega})$ in order to arrive at a final solution through the inverse transform. Usually the function $F(\underset{\sim}{\omega})$ is assumed to be either even or odd.

Weighting functions corresponding to several common one-dimensional covariance functions have been determined by a number of authors, notably Journel and Huijbregts (1978) and Mantoglou and Wilson (1981). In higher dimensions, the calculation of weighting functions becomes quite complex and is often done numerically

using FFT's. The non-uniqueness of the weighting function and the difficulty in finding it, particularly in higher dimensions, renders this method of questionable value to the user who wishes to be able to handle arbitrary covariance functions.

Leaving this issue for the moment, the implementation of the MA method is itself a rather delicate problem. For a discrete process in one dimension, Eq. can be written

$$Z_i = \sum_{j=-\infty}^{\infty} f_j \, W_{i,j}, \tag{23}$$

where $W_{i,j}$ is a discrete white noise process taken to have zero mean and unit variance. To implement this in practice, the sum must be restricted to some range p, usually chosen such that $f_{\pm p}$ is negligible,

$$z_i = \sum_{j=-p}^{p} f_j \, W_{i,j}. \tag{24}$$

The next concern is how to discretize the underlying white noise process. If Δx is the increment of the physical process such that $z_i = z((i-1)\Delta x)$ and Δu is the incremental distance between points of the underlying white noise process, such that

$$W_{i,j} = W((i-1)\Delta x + j\Delta u), \tag{25}$$

then $f_j = f(j\Delta u)$ and Δu should be chosen such that the quotient $r = \Delta x/\Delta u$ is an integer for simplicity. Figure 2 illustrates the relationship between Z_i and the discrete white noise process. For finite Δu, the discrete approximation (24) will introduce some error into the estimated covariance of the realization. This error can often be removed through a multiplicative correction factor as shown by Journel and Huijbregts (1978) but in general is reduced by taking Δu as small as practically possible (and thus p as large as possible).

Once the discretization of the underlying white noise process and the range p has been determined, the implementation of (24) in one dimension is quite straightforward and usually quite efficient for reasonable values of p. In higher dimensions, the method rapidly becomes cumbersome. Figure 3 shows a typical portion of a 2-D discrete process Z_{ij}, marked by X's, and the underlying white noise field, marked by dots. The entire figure represents the upper right corner of a 2-D field. The process z_{ij} is now formed by the double summation

$$z_{ij} = \sum_{k=-p_1}^{p_1} \sum_{\ell=-p_2}^{p_2} f_{k\ell} \, W_{i,j,k,\ell}, \tag{26}$$

where $f_{k\ell}$ is the 2-D weighting function and $W_{i,j,k,\ell}$ is the discrete white noise process centered at the same position as z_{ij}. The i and j subscripts on W are for bookkeeping purposes so that the sum is performed over a centered neighborhood of discrete white noise values.

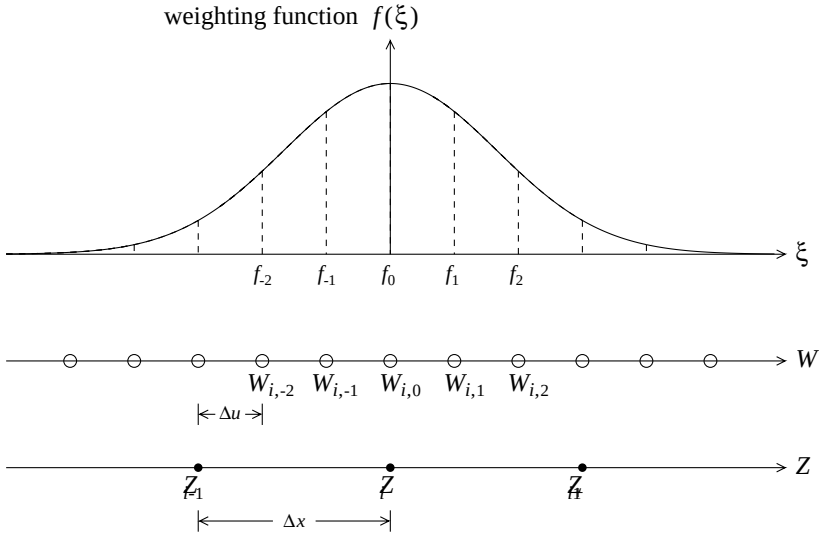


Figure 2. Schematic representation of the moving average process in one dimension.

In the typical example illustrated in Figure 3, the discretization of the white noise process is such that $r = \Delta u/\Delta x = 3$ and a relatively short scale of fluctuation was used so that $p = 6$. This means that if a $K_1 \times K_2$ field is to be simulated, the total number of white noise realizations to be generated must be,

$$N_W = \Big(1 + 2p_1 + r_1(K_1 - 1)\Big)\Big(1 + 2p_2 + r_2(K_2 - 1)\Big), \tag{27}$$

or about $(rK)^2$ for a square field. This can be contrasted immediately with the FFT approach which requires the generation of about $\frac{1}{2}K^2$ random values for a quadrant symmetric process (note that the factor of one-half is a consequence of the periodicity of the generated field). When $r = 3$, some 18 times as many white noise realizations must be generated for the moving average algorithm as for the FFT method. Also the construction of each field point requires a total of $(2p + 1)^2$ additions and multiplications which, for the not unreasonable example given above, is $13^2 = 169$. This means that the entire field will be generated using $K^2(2p + 1)^2$ or about 11 million additions and multiplications for a 200×200 field. Again this can be contrasted to the two-dimensional FFT method (radix-2, row-column algorithm) which requires some $4K^2 \log_2 K$ or about 2 million multiply-adds. In most cases, the moving average approach in two dimensions was found to run at least 10 times slower than the FFT approach. In three dimensions, the moving average method used to generate a $64 \times 64 \times 64$ field with $p = 6$ was estimated to run over 100 times slower than the corresponding FFT approach. For this reason, and since the weighting function is generally difficult to find, the moving average method as a general method of producing realizations of multi-dimensional random fields is only useful when the moving average representation is particularly desired.

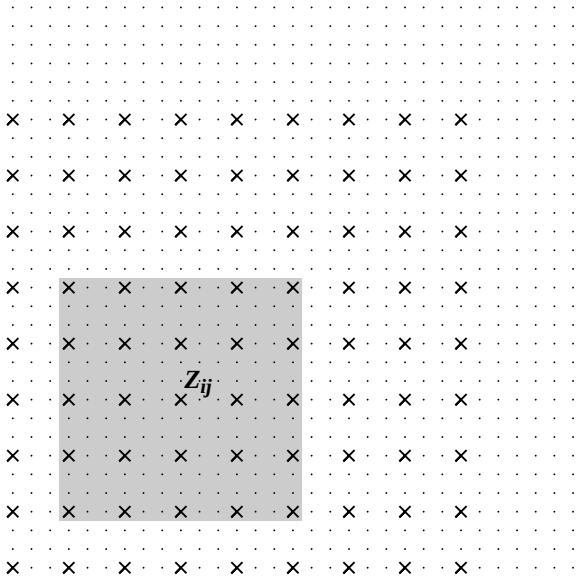


Figure 3. Two-dimensional moving average process. Z_{ij} is formed by summing the contributions from the underlying white noise process in the shaded region.

It can be noted in passing that the two-dimensional ARMA model suggested by Naganum et al (1987) requires about 50 to 150 multiply-adds (depending on the type of covariance structure modeled) for each field point. This is about 2 to 6 times slower than the FFT approach. While this is quite competitive for certain covariance functions, the corresponding run speeds for three-dimensional processes are estimated to be 15 to 80 times slower than the FFT approach depending on the choice of parameters p and r. Also, in a sequence of two papers, Mignolet and Spanos (1992) and Spanos and Mignolet (1992) discuss in considerable detail the moving average (MA), autoregressive (AR) and ARMA approaches to simulating two-dimensional random fields. In their examples, they obtain accurate results at the expense of running about 10 or more times slower than the fastest of the methods to be considered later in this chapter.

4.2 Covariance Matrix Decomposition

Covariance matrix decomposition is a direct method of producing a homogeneous random field with prescribed covariance structure $C(\underset{\sim}{x}_i - \underset{\sim}{x}_j) = C(\underset{\sim}{\tau}_{ij})$, where $\underset{\sim}{x}_i$, $i = 1, 2, \ldots, n$ are discrete points in the field and $\underset{\sim}{\tau}_{ij}$ is the lag vector between the points $\underset{\sim}{x}_i$ and $\underset{\sim}{x}_j$. If $\underset{\approx}{C}$ is a positive definite covariance matrix with elements $C_{ij} = C(\underset{\sim}{\tau}_{ij})$, then a mean zero discrete process $Z_i = Z(\underset{\sim}{x}_i)$ can be produced (using vector notation) according to

$$\underset{\sim}{Z} = \underset{\approx}{L}\underset{\sim}{U} \tag{28}$$

where $\underset{\approx}{L}$ is a lower triangular matrix satisfying $\underset{\approx}{L}\underset{\approx}{L}^T = \underset{\approx}{C}$ (typically obtained using Cholesky decomposition) and $\underset{\sim}{U}$ is a vector of n independent mean zero, unit variance Gaussian random variables. Although appealing in its simplicity and accuracy, this method is only useful for small fields. In two dimensions, the covariance matrix of a 128×128 field would be of size $16,384 \times 16,384$ and the Cholesky decomposition of such a matrix would be both time consuming and prone to considerable round-off error (particularly since covariance matrices are often numerically singular).

4.3 Discrete Fourier Transform Method

The Fourier Transform method is based on the spectral representation of homogeneous mean square continuous random fields, $Z(\underset{\sim}{x})$, which can be expressed as follows (Yaglom, 1962)

$$Z(\underset{\sim}{x}) = \int_{-\infty}^{\infty} e^{i\underset{\sim}{x}\cdot\underset{\sim}{\omega}} \, W(d\underset{\sim}{\omega}) \tag{29}$$

where $W(d\underset{\sim}{\omega})$ is an interval white noise process with mean zero and variance $S(\underset{\sim}{\omega})\, d\underset{\sim}{\omega}$. This representation is in terms of the physically meaningful spectral density function, $S(\underset{\sim}{\omega})$, and so is intuitively attractive. In practice, the n-dimensional integral becomes an n-dimensional sum which is evaluated separately at each point $\underset{\sim}{x}$. Although potentially accurate, the method is computationally slow for reasonable field sizes and typical spectral density functions – the DFT is generally about as efficient as the MA discussed above. Its major advantage over the MA approach is that the spectral density function is estimated in practice using standard techniques.

In n-dimensions, for real $Z(\underset{\sim}{x})$, the Discrete Fourier Transform (DFT) can be written

$$Z(\underset{\sim}{x}) = \sum_{k_1=-N_1}^{N_1} \cdots \sum_{k_n=-N_n}^{N_n} C_{k_1 \ldots k_n} \cos\left(\omega_{k_1} x_1 + \cdots \omega_{k_n} x_n + \Phi_{k_1 \ldots k_n}\right) \tag{30}$$

where $\Phi_{k_1 \ldots k_n}$ is a random phase angle uniformly distributed on $[0, 2\pi]$ and $C_{k_1 k_2 \ldots k_n}$ is a random amplitude having Rayleigh distribution if Z is Gaussian. An alternative way of writing the DFT is

$$\begin{aligned} Z(\underset{\sim}{x}) = \sum_{k_1=-N_1}^{N_1} \cdots \sum_{k_n=-N_n}^{N_n} & A_{k_1 \ldots k_n} \cos\left(\omega_{k_1} x_1 + \cdots \omega_{k_n} x_n\right) \\ & + B_{k_1 \ldots k_n} \sin\left(\omega_{k_1} x_1 + \cdots \omega_{k_n} x_n\right) \end{aligned} \tag{31}$$

where, for a stationary normally distributed $Z(\underset{\sim}{x})$, the A and B coefficients are mutually independent and normally distributed with zero means and variances given by

$$\mathrm{E}\left[A^2_{k_1 k_2 \ldots k_n}\right] = \mathrm{E}\left[B^2_{k_1 k_2 \ldots k_n}\right] = S(\underset{\sim}{\omega}_{\underset{\sim}{k}})\Delta\underset{\sim}{\omega} \tag{32}$$

In this equation, $\underset{\sim}{\omega}_{\underset{\sim}{k}} = \{\omega_{k_1}, \omega_{k_2}, \ldots, \omega_{k_n}\}$ and $S(\underset{\sim}{\omega}_{\underset{\sim}{k}}\Delta\underset{\sim}{\omega}$ is the area under the spectral density function in an incremental region centered on $\underset{\sim}{\omega}_{\underset{\sim}{k}}$.

As mentioned above, the sum is composed of $(2N+1)^n$ terms (if $N_1 = N_2 = \cdots = N$), where $(2N+1)$ is the number of discrete frequencies taken in each dimension. Depending on the shape of the spectral density function, N might easily be of the order of 100, so that in 3 dimensions roughly 8 million terms must be summed for each spatial position desired in the generated field (thus, in 3 dimensions, a 20 x 20 x 20 random field would involve roughly 128 billion evaluations of sin or cosine).

This approach is really only computationally practical in one-dimension where the DFT reduces to

$$Z(x) = \sum_{k=-N}^{N} A_k \cos(\omega_k x) + B_k \sin(\omega_k x) \tag{33}$$

where

$$\begin{aligned} \mathrm{E}\big[A_k\big] &= \mathrm{E}\big[B_k\big] = 0 \\ \mathrm{E}\big[A_k^2\big] &= \mathrm{E}\big[B_k^2\big] = S(\omega_k)\Delta\omega \end{aligned} \tag{34}$$

and where the A and B coefficients are mutually independent of all other A's and B's. If the symmetry in the spectral density function is taken advantage of, namely that $S(\omega) = S(-\omega)$, then the sum can be written

$$Z(x) = \sum_{k=0}^{N} A_k \cos(\omega_k x) + B_k \sin(\omega_k x) \tag{35}$$

where now the variances of the A and B coefficients are expressed in terms of the one-sided spectral density function

$$\mathrm{E}\big[A_k^2\big] = \mathrm{E}\big[B_k^2\big] = G(\omega_k)\Delta\omega_k \tag{36}$$

and where $\Delta\omega_0 = \frac{1}{2}(\omega_1 - \omega_0)$ and $\Delta\omega_k = \frac{1}{2}(\omega_{k+1} - \omega_{k-1})$.

Simulation proceeds as follows;

1. decide on how to discretize the spectral density (ie. on N and $\Delta\omega$),
2. generate mean zero, normally distributed, realizations of A_k and B_k for $k = 0, 1, \ldots, N$ each having variance given by Eq. 36
3. for each value of x desired in the final random process, compute the sum given by Eq. 35.

4.4 The Fast Fourier Transform Method

If both space and frequency are discretized into a series of equispaced points, then the *Fast Fourier Transform* (FFT) method developed by Cooley and Tukey (1965) can be used to compute Eq. 29. The FFT is much more computationally efficient than the DFT. For example, in one-dimension the DFT requires N^2 operations whereas the

FFT requires only $N \log_2 N$ operations. If $N = 2^{15} = 32768$, then the FFT will be approximately 2000 times faster than the DFT. For the purposes of this development, only the one-dimensional case will be considered and multi-dimensional results will be stated subsequently. For real and discrete $Z(x_j)$, $j = 1, 2, \ldots, N$, Eq. 29 becomes

$$Z(x_j) = \int_{-\pi}^{\pi} e^{ix_j\omega}\, W(d\omega) \;=\; \lim_{K\to\infty} \sum_{k=-K}^{K} e^{ix_j\omega_k}\, W(\Delta\omega_k)$$

$$= \lim_{K\to\infty} \sum_{k=-K}^{K} \Big\{ A(\Delta\omega_k)\cos(x_j\omega_k) + B(\Delta\omega_k)\sin(x_j\omega_k) \Big\} \tag{37}$$

where $\omega_k = k\pi/K$, $\Delta\omega_k$ is an interval of length π/K centered at ω_k, and the last step in (37) follows from the fact that Z is real. The functions $A(\Delta\omega_k)$ and $B(\Delta\omega_k)$ are independent identically distributed random interval functions with mean zero and $\mathrm{E}\big[A(\Delta\omega_k)A(\Delta\omega_m)\big] = \mathrm{E}\big[B(\Delta\omega_k)B(\Delta\omega_m)\big] = 0$ for all $k \neq m$ in the limit as $\Delta\omega \to 0$. At this point, the simulation involves generating realizations of $A_k = A(\Delta\omega_k)$ and $B_k = B(\Delta\omega_k)$ and evaluating (37). Since the process is real, $S(\omega) = S(-\omega)$, and the variances of A_k and B_k can be expressed in terms of the one-sided spectral density function $G(\omega) = 2S(\omega)$, $\omega \geq 0$. This means that the sum in (37) can have lower bound $k = 0$. Note that an equivalent way of writing (37) is

$$Z(x_j) = \sum_{k=0}^{K} C_k \cos(x_j\omega_k + \Phi_k), \tag{38}$$

where Φ_k is a random phase angle uniformly distributed on $[0, 2\pi]$ and C_k follows a Rayleigh distribution. Shinozuka and Jan (1972) take $C_k = \sqrt{G(\omega_k)\Delta\omega}$ to be deterministic, an approach not followed here since it gives an upper bound on Z over the space of outcomes of $Z \leq \sum_{k=0}^{K} \sqrt{G(\omega_k)\Delta\omega}$ which may be an unrealistic restriction, particularly in reliability calculations which could very well depend on extremes.

Next, the process $Z_j = Z(x_j)$ is assumed to be periodic, $Z_j = Z_{K+j}$, with the same number of spatial and frequency discretization points ($N = K$). As will be shown later, the periodicity assumption leads to a symmetric covariance structure which is perhaps the major disadvantage to the DFT and FFT approach. If the physical length of the one-dimensional process under consideration is D and the space and frequency domains are discretized according to

$$x_j = j\Delta x \;=\; \frac{jD}{K-1} \tag{39}$$

$$\omega_j = j\Delta\omega \;=\; \frac{2\pi j(K-1)}{KD} \tag{40}$$

for $j = 0, 1, \ldots, K-1$, then the Fourier transform

$$Z_j = \sum_{k=0}^{K-1} \mathcal{X}_k\, e^{i(2\pi jk/K)} \tag{41}$$

can be evaluated using the FFT algorithm. The Fourier coefficients, $\mathcal{X}_k = A_k - iB_k$, have the following symmetries due to the fact that Z is real,

$$A_k = \frac{1}{K}\sum_{j=0}^{K-1} Z_j \cos 2\pi\tfrac{jk}{K} \;=\; A_{K-k} \tag{42}$$

$$B_k = \frac{1}{K}\sum_{j=0}^{K-1} Z_j \sin 2\pi\tfrac{jk}{K} \;=\; -B_{K-k} \tag{43}$$

which means that A_k and B_k need only be generated randomly for $k = 0, 1, \ldots, K/2$ and that $B_0 = B_{K/2} = 0$. Note that if the coefficients at $K-k$ are produced independently of the coefficients at k, the resulting field will display aliasing. Thus there is no advantage to taking Z to be complex, generating all the Fourier coefficients randomly, and attempting to produce two independent fields simultaneously (the real and imaginary parts), or in just ignoring the imaginary part.

As far as the simulation is concerned, all that remains is to specify the statistics of A_k and B_k so that they can be generated randomly. If Z is a Gaussian mean zero process, then so are A_k and B_k. The variance of A_k can be computed in a consistent fashion by evaluating $\mathrm{E}\big[A_k^2\big]$ using (42)

$$\mathrm{E}\big[A_k^2\big] = \frac{1}{K^2}\sum_{j=0}^{K-1}\sum_{\ell=0}^{K-1} \mathrm{E}\big[Z_j\, Z_\ell\big] \cos 2\pi\tfrac{jk}{K} \cos 2\pi\tfrac{\ell k}{K} \tag{44}$$

This result suggests using the covariance function directly to evaluate the variance of A_k, however the implementation is complex and no particular advantage in accuracy is attained. A simpler approach involves the discrete approximation to the Wiener-Khinchine relationship

$$\mathrm{E}\big[Z_j\, Z_\ell\big] \simeq \Delta\omega \sum_{m=0}^{K-1} G(\omega_m) \cos 2\pi\tfrac{m(j-\ell)}{K} \tag{45}$$

which when substituted into (44) leads to

$$\begin{aligned}
\mathrm{E}\big[A_k^2\big] &= \frac{\Delta\omega}{K^2}\sum_{j=0}^{K-1}\sum_{\ell=0}^{K-1}\sum_{m=0}^{K-1} G(\omega_m) \cos 2\pi\tfrac{m(j-\ell)}{K} C_{kj}C_{k\ell} \\
&= \frac{\Delta\omega}{K^2}\sum_{m=0}^{K-1} G(\omega_m) \sum_{j=0}^{K-1} C_{mj}C_{kj} \sum_{\ell=0}^{K-1} C_{m\ell}C_{k\ell} \\
&+ \frac{\Delta\omega}{K^2}\sum_{m=0}^{K-1} G(\omega_m) \sum_{j=0}^{K-1} S_{mj}C_{kj} \sum_{\ell=0}^{K-1} S_{m\ell}C_{k\ell},
\end{aligned} \tag{46}$$

where $C_{kj} = \cos 2\pi\frac{kj}{K}$ and $S_{kj} = \sin 2\pi\frac{kj}{K}$.

To reduce (46) further, use is made of the following two identities

1) $$\sum_{k=0}^{K-1} \sin 2\pi \tfrac{mk}{K} \cos 2\pi \tfrac{jk}{K} = 0$$

2) $$\sum_{k=0}^{K-1} \cos 2\pi \tfrac{mk}{K} \cos 2\pi \tfrac{jk}{K} = \begin{cases} 0, & \text{if } m \neq j \\ \frac{K}{2}, & \text{if } m = j \text{ or } K-j \\ K, & \text{if } m = j = 0 \text{ or } \frac{K}{2} \end{cases}$$

By identity (1), the second term of (46) is zero. The first term is also zero, except when $m = k$ or $m = K - k$, leading to the results

$$\mathrm{E}\big[A_k^2\big] = \begin{cases} \frac{1}{2}G(\omega_k)\Delta\omega, & \text{if } k = 0 \\ \frac{1}{4}\big\{G(\omega_k) + G(\omega_{K-k})\big\}\Delta\omega, & \text{if } k = 1, \ldots, \frac{K}{2} - 1 \\ G(\omega_k)\Delta\omega, & \text{if } k = \frac{K}{2} \end{cases} \quad (47)$$

remembering that for $k = 0$ the frequency interval is $\frac{1}{2}\Delta\omega$. An entirely similar calculation leads to

$$\mathrm{E}\big[B_k\big]^2 = \begin{cases} 0, & \text{if } k = 0 \text{ or } \frac{K}{2} \\ \frac{1}{4}\big\{G(\omega_k) + G(\omega_{K-k})\big\}\Delta\omega, & \text{if } k = 1, \ldots, \frac{K}{2} - 1 \end{cases} \quad (48)$$

Thus the simulation process is as follows;

1. generate independent normally distributed realizations of A_k and B_k having mean zero and variance given by (47) and (48) for $k = 0, 1, \ldots, K/2$ and set $B_0 = B_{K/2} = 0$,
2. use the symmetry relationships, (42) and (43), to produce the remaining Fourier coefficients for $k = 1 + K/2, \ldots, K - 1$
3. produce the field realization by Fast Fourier Transform using Eq. 41.

In higher dimensions a similar approach can be taken. To compute the Fourier sum over non-negative frequencies only, the spectral density function $S(\underset{\sim}{\omega})$ is assumed to be even in all components of $\underset{\sim}{\omega}$ (quadrant symmetric) so that the 'one-sided' spectral density function, $G(\underset{\sim}{\omega}) = 2^n S(\underset{\sim}{\omega}) \; \forall \; \omega_i \geq 0$, and n-dimensional space, can be employed. Using $L = K_1 - \ell$, $M = K_2 - m$, and $N = K_3 - n$ to denote the symmetric points in fields of size $K_1 \times K_2$ in 2-D or $K_1 \times K_2 \times K_3$ in 3-D, the Fourier coefficients yielding a real two-dimensional process must satisfy

$$\begin{aligned} A_{LM} &= A_{\ell m}, & B_{LM} &= -B_{\ell m} \\ A_{\ell M} &= A_{Lm}, & B_{\ell M} &= -B_{Lm} \end{aligned} \quad (49)$$

for $\ell, m = 0, 1, \ldots, \frac{K_\alpha}{2}$ where K_α is either K_1 or K_2 appropriately. Note that these relationships are applied modulo K_α, so that $A_{K_1-0,m} \equiv A_{0,m}$ for example. In two dimensions, the Fourier coefficients must be generated over two adjacent quadrants of

the field, the rest of the coefficients obtained using the symmetry relations. In three dimensions, the symmetry relationships are

$$\begin{aligned} A_{LMN} &= A_{\ell mn}, & B_{LMN} &= -B_{\ell mn} \\ A_{\ell MN} &= A_{Lmn}, & B_{\ell MN} &= -B_{Lmn} \\ A_{LmN} &= A_{\ell Mn}, & B_{LmN} &= -B_{\ell Mn} \\ A_{\ell mN} &= A_{LMn}, & B_{\ell mN} &= -B_{LMn} \end{aligned} \tag{50}$$

for $\ell, m, n = 0, 1, \ldots, \frac{K_\alpha}{2}$. Again, only half the Fourier coefficients are to be generated randomly.

The variances of the Fourier coefficients are found in a manner analogous to the one-dimensional case, resulting in

$$\mathrm{E}\big[A_{\ell m}^2\big] = \tfrac{1}{8}\delta_{\ell m}^A \Delta\underset{\sim}{\omega}\Big(G_{\ell m}^d + G_{\ell N}^d + G_{Ln}^d + G_{LN}^d\Big) \tag{51}$$

$$\mathrm{E}\big[B_{\ell m}^2\big] = \tfrac{1}{8}\delta_{\ell m}^B \Delta\underset{\sim}{\omega}\Big(G_{\ell m}^d + G_{\ell N}^d + G_{Ln}^d + G_{LN}^d\Big) \tag{52}$$

for two-dimensions and

$$\begin{aligned} \mathrm{E}\big[A_{\ell mn}\big]^2 = \tfrac{1}{16}\,\delta_{\ell mn}^A \Delta\underset{\sim}{\omega}\,\Big(&G_{\ell mn}^d + G_{\ell mN}^d + G_{\ell Mn}^d + G_{Lmn}^d \\ &+ G_{\ell MN}^d + G_{LmN}^d + G_{LMn}^d + G_{LMN}^d\Big) \end{aligned} \tag{53}$$

$$\begin{aligned} \mathrm{E}\big[B_{\ell mn}\big]^2 = \tfrac{1}{16}\,\delta_{\ell mn}^B \Delta\underset{\sim}{\omega}\,\Big(&G_{\ell mn}^d + G_{\ell mN}^d + G_{\ell Mn}^d + G_{Lmn}^d \\ &+ G_{\ell MN}^d + G_{LmN}^d + G_{LMn}^d + G_{LMN}^d\Big) \end{aligned} \tag{54}$$

in three-dimensions, where for p dimensions,

$$\Delta\underset{\sim}{\omega} = \prod_{i=1}^{p} \Delta\omega_i, \tag{55}$$

$$G^d(\underset{\sim}{\omega}) = \frac{G(\omega_1, \ldots, \omega_p)}{2^d}, \tag{56}$$

and d is the number of components of $\underset{\sim}{\omega} = (\omega_1, \ldots, \omega_p)$ which are equal to zero. The factors $\delta_{\ell mn}^A$ and $\delta_{\ell mn}^B$ are given by

$$\delta_{\ell mn}^A = \begin{cases} 2 & \text{if } \ell = 0 \text{ or } \frac{K_1}{2} \text{ and } m = 0 \text{ or } \frac{K_2}{2} \text{ and } n = 0 \text{ or } \frac{K_3}{2} \\ 1 & \text{otherwise} \end{cases} \tag{57}$$

$$\delta_{\ell mn}^B = \begin{cases} 0 & \text{if } \ell = 0 \text{ or } \frac{K_1}{2} \text{ and } m = 0 \text{ or } \frac{K_2}{2} \text{ and } n = 0 \text{ or } \frac{K_3}{2} \\ 1 & \text{otherwise} \end{cases} \tag{58}$$

(ignoring the index n in the case of two dimensions). Thus, in higher dimensions, the simulation procedure is almost identical to that followed in the 1-D case – the only difference being that the coefficients are generated randomly over the half plane (2-D) or the half volume (3-D) rather than the half line of the 1-D formulation.

It is appropriate at this time to investigate some of the shortcomings of the method. First of all it is easy to show that regardless of the desired target covariance function, the covariance function $\hat{C}_k = \hat{C}(k\Delta x)$ of the real FFT process is always symmetric about the midpoint of the field. In one-dimension, the covariance function is given by (using complex notation for the time being),

$$\begin{aligned}\hat{C}_k &= \mathrm{E}\big[Z_{\ell+k}\overline{Z_\ell}\,\big] \\ &= \mathrm{E}\left[\sum_{j=0}^{K-1} \mathcal{X}_j \exp\left\{i\left(\tfrac{2\pi(\ell+k)j}{K}\right)\right\} \sum_{m=0}^{K-1} \overline{\mathcal{X}_m} \exp\left\{-i\left(\tfrac{2\pi\ell m}{K}\right)\right\}\right] \\ &= \sum_{j=0}^{K-1} \mathrm{E}\big[\mathcal{X}_j\overline{\mathcal{X}_j}\,\big] \exp\left\{i\left(\tfrac{2\pi jk}{K}\right)\right\}, \end{aligned} \tag{59}$$

where use was made of the fact that $\mathrm{E}\big[\mathcal{X}_j\overline{\mathcal{X}_m}\,\big] = 0$ for $j \neq m$ (overbar denotes the complex conjugate). Similarly one can derive

$$\hat{C}_{K-k} = \sum_{j=0}^{K-1} \mathrm{E}\big[\mathcal{X}_j\overline{\mathcal{X}_j}\,\big] \exp\left\{-i\left(\tfrac{2\pi jk}{K}\right)\right\} \;=\; \overline{\hat{C}_k} \tag{60}$$

since $\mathrm{E}\big[\mathcal{X}_j\overline{\mathcal{X}_j}\,\big]$ is real. The covariance function of a real process is also real in which case (60) becomes simply

$$\hat{C}_{K-k} = \hat{C}_k. \tag{61}$$

In one dimension, this symmetry is illustrated by Figure 4. Similar results are observed in higher dimensions. In general, this deficiency can be overcome by generating a field twice as long as required in each coordinate direction and keeping only the first quadrant of the field. Figure 4 also compares the covariance, mean, and variance fields of the LAS method to that of the FFT method (the TBM method is not defined in one dimension). The two methods give satisfactory performance with respect to the variance and mean fields, while the LAS method shows superior performance with respect to the covariance structure.

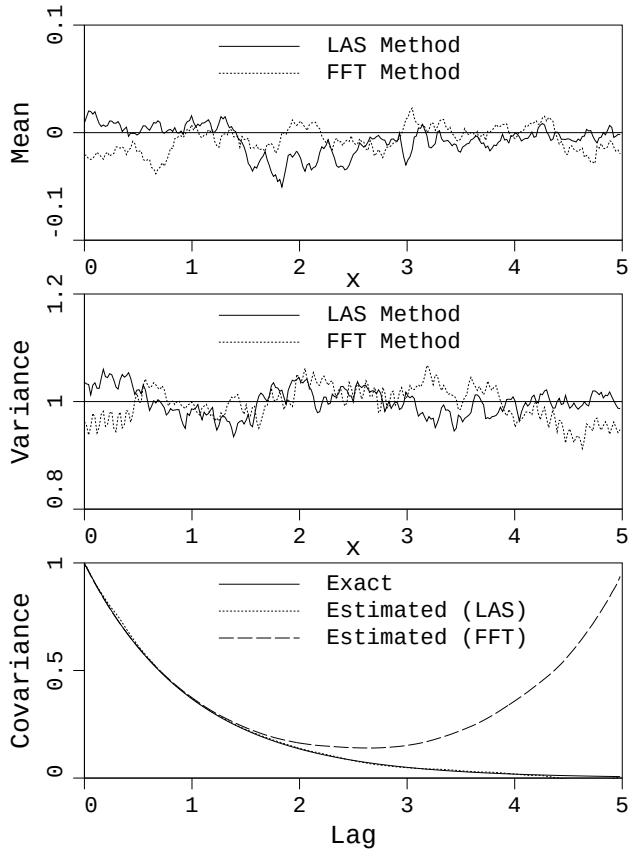
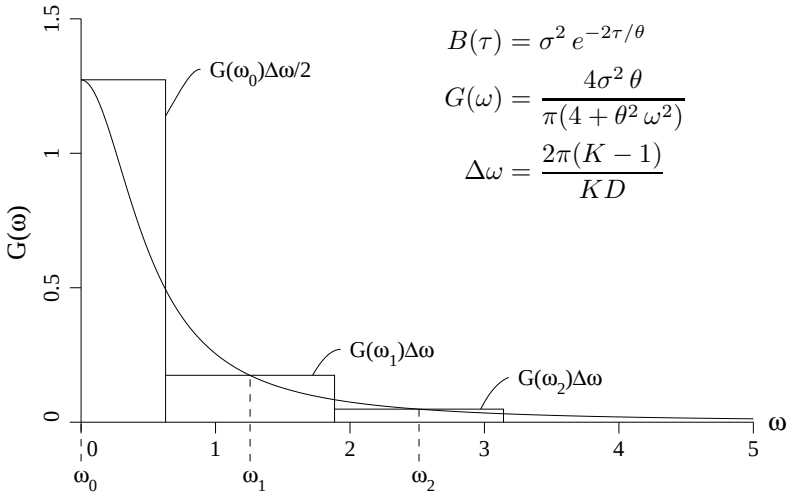


Figure 4. Mean, variance, and covariance of a 1-D 128 point Gauss-Markov process estimated over an ensemble of 2000 realizations generated by the FFT method.

The second problem with the FFT method relates primarily to its ease of use. Because of the close relationship between the spatial and frequency discretization, considerable care must be exercised when initially defining the spatial field and its discretization. First of all the physical length of the field D must be large enough that the frequency increment $\Delta\omega = 2\pi(K-1)/KD \simeq 2\pi/D$ is sufficiently small. This is necessary if the sequence $\frac{1}{2}G(\omega_0)\Delta\omega, G(\omega_1)\Delta\omega, \ldots$ is to adequately approximate the target spectral density function. Figure 5 shows an example where the frequency discretization is overly coarse. Secondly, the physical resolution Δx must be selected so that the spectral density above the frequency $2\pi/\Delta x$ is negligible. Failure to do so will result in an underestimation of the total variance of the process. In fact the FFT formulation given above folds the power corresponding to frequencies between $\pi/\Delta x$ and $2\pi/\Delta x$ into the power at frequencies below the Nyquist limit $\pi/\Delta x$. This results in the point variance of the simulation being more accurate than if the power above the Nyquist limit were ignored, however it leads to a non-uniqueness in that a family of spectral

density functions, all having the same value of $G(\omega_k) + G(\omega_{K-k})$, yield the same process. In general it is best to choose Δx so that the power above the Nyquist limit is negligible. The second term involving the symmetric frequency $G(\omega_{K-k})$ is included here because the point variance is the most important second-order characteristic.

Unfortunately, many applications dictate the size and discretization of the field *a-priori* or the user may want to have the freedom to easily consider other geometries or spectral density functions. Without a good deal of careful thought and analysis, the FFT approach can easily yield highly erroneous results.



$$B(\tau) = \sigma^2 \, e^{-2\tau/\theta} \tag{62}$$

$$G(\omega) = \frac{4\sigma^2 \, \theta}{\pi(4 + \theta^2 \, \omega^2)}$$

$$\Delta\omega = \frac{2\pi(K-1)}{KD} \tag{63}$$

Figure 5. Example of overly coarse frequency discretization resulting in a poor estimation of point variance ($D = 5$ and $\theta = 4$).

A major advantage of the FFT method is that it can easily handle anisotropic fields with no sacrifice in efficiency. The field need not be square, although many implementations of the FFT require the number of points in the field in any coordinate direction to be a power of two. Regarding efficiency, it should be pointed out that the time to generate the first realization of the field is generally much longer than that required to generate subsequent realizations. This is because the statistics of the Fourier coefficients must be calculated only once (see Eq.'s 47 and 48).

The FFT method is useful for the generation of fractal processes, which are most naturally represented by the spectral density function. In fact the covariance function does not exist since the variance of a fractal process is ideally infinite. In practice, for such a process, the spectral density is truncated above and below to render a finite variance realization.

4.5 The Turning Bands Method

The Turning Bands Method (TBM), as originally suggested by Matheron (1973), involves the simulation of random fields in two- or higher-dimensional space by using a sequence of one-dimensional processes along lines crossing the domain. With reference to Figure 6, the algorithm can be described as follows,

1. choose an arbitrary origin within or near the domain of the field to be generated,
2. select a line i crossing the domain having a direction given by the unit vector $\underset{\sim}{u}_i$ which may be chosen either randomly or from some fixed set,
3. generate a realization of a one-dimensional process, $Z_i(\xi_i)$, along the line i having zero mean and covariance function $B_1(\tau_i)$ where ξ_i and τ_i are measured along line i,
4. orthogonally project each field point $\underset{\sim}{x}_k$ onto the line i to define the coordinate ξ_{ki} ($\xi_{ki} = \underset{\sim}{x}_k \cdot \underset{\sim}{u}_i$ in the case of a common origin) of the one-dimensional process value $Z_i(\xi_{ki})$,
5. add the component $Z_i(\xi_{ki})$ to the field value $Z(\underset{\sim}{x}_k)$ for each $\underset{\sim}{x}_k$,
6. return to step (2) and generate a new one-dimensional process along a subsequent line until L lines have been produced,
7. normalize the field $Z(\underset{\sim}{x}_k)$ by dividing through by the factor $\sqrt{L}$.

Essentially, the generating equation for the zero-mean discrete process $Z(\underset{\sim}{x})$ is given by

$$Z(\underset{\sim}{x}_k) = \frac{1}{\sqrt{L}} \sum_{i=1}^{L} Z_i(\underset{\sim}{x}_k \cdot \underset{\sim}{u}_i), \tag{64}$$

where if the origins of the lines and space are not common, the dot product must be replaced by some suitable transform. This formulation depends on knowledge of the one-dimensional covariance function, $B_1(\tau)$. Once this is known, the line processes can be produced using some efficient 1-D algorithm.

The covariance function $B_1(\tau)$ is chosen such that the multi-dimensional covariance structure $B_n(\underset{\sim}{\tau})$ in n-dimensional space is reflected over the ensemble. For two-dimensional isotropic processes, Mantoglou and Wilson (1981) give the following relationship between $B_2(\underset{\sim}{\tau})$ and $B_1(\eta)$ for $r = |\underset{\sim}{\tau}|$,

$$B_2(r) = \frac{2}{\pi} \int_0^r \frac{B_1(\eta)}{\sqrt{r^2 - \eta^2}}\, d\eta, \tag{65}$$

which is an integral equation to be solved for $B_1(\eta)$. In three dimensions, the relationship between the isotropic $B_3(r)$ and $B_1(\eta)$ is particularly simple,

$$B_1(\eta) = \frac{d}{d\eta}\Big(\eta\, B_3(\eta)\Big). \tag{66}$$

Mantoglou and Wilson supply explicit solutions for either the equivalent one-dimensional covariance function or the equivalent one-dimensional spectral density function for a variety of common multi-dimensional covariance structures.

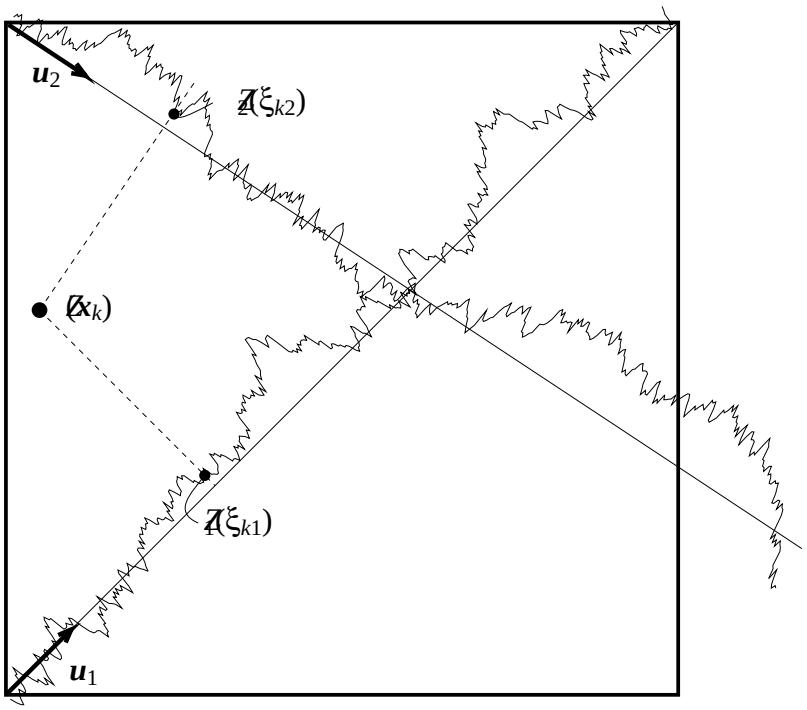


Figure 6. The Turning Bands Method: contributions from the line process $Z_i(\xi_i)$ at the closest point are summed into the field process $Z(\underset{\sim}{x})$ at $\underset{\sim}{x}_k$.

In this implementation of the TBM, the line processes were constructed using a 1-D FFT algorithm as discussed in the previous section. The LAS method was not used for this purpose because the local averaging introduced by the method would complicate the resulting covariance function of (65). Line lengths were chosen to be twice that of the field diagonal to avoid the symmetric covariance problem inherent with the FFT method. To reduce errors arising due to overly coarse discretization of the lines, the ratio between the incremental distance along the lines, $\Delta\xi$, and the minimum incremental distance in the field along any coordinate, Δx, was selected to be $\Delta\xi/\Delta x = \frac{1}{2}$.

Figure 7 represents a realization of a 2-D process. The finite number of lines used, in this case 16, results in a streaked appearance of the realization. A number of origin locations were experimented with to mitigate the streaking, the best appearing to be the use of all four corners as illustrated in Figure 6 and as used in Figure 7. The corner selected as an origin depends on which quadrant the unit vector $\underset{\sim}{u}_i$ points into. If one considers the spectral representation of the one-dimensional random processes along

each line (see 29) it is apparent that the streaks are a result of constructive/destructive interference between randomly oriented traveling plane waves. The effect will be more pronounced for narrow band processes and for a small number of lines. For this particular covariance function (Markov), the streaks are still visible when 32 lines are used, but, as shown in Figure 8, are negligible when using 64 lines (the use of number of lines which are powers of 2 is arbitrary). While the 16 line case runs at about the same speed as the 2-D LAS approach, the elimination of the streaks in the realization comes at a price of running about 4 times slower. The streaks are only evident in an average over the ensemble if non-random line orientations are used, although they still appear in individual realizations in either case. Thus, with respect to each realization, there is no particular advantage to using random versus non-random line orientations.

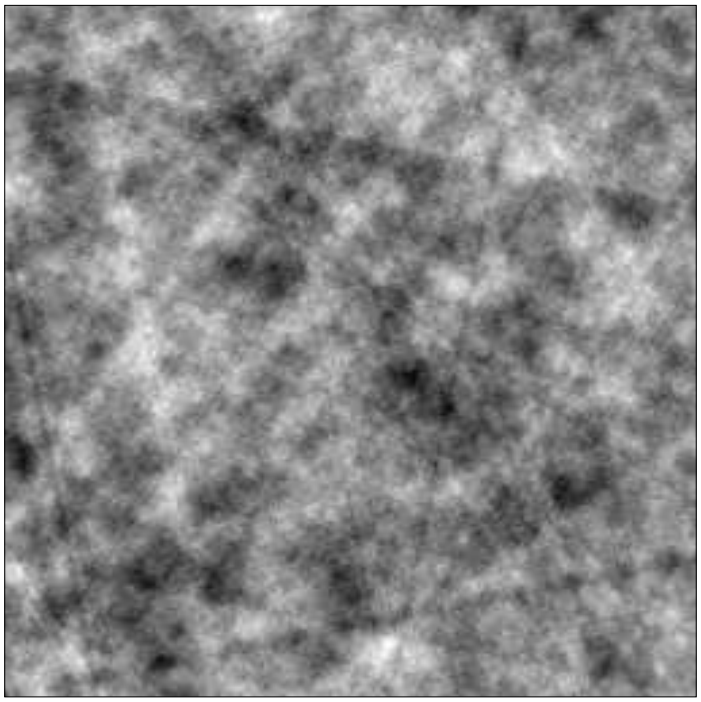


Figure 7. Sample function of a 2-D field via TBM using 16 lines.

Since the streaks are present in the field itself, this type of error is generally more serious than errors in the variance or covariance field. For example, if the field is being used to represent soil conductivity, then the streaks could represent paths of reduced resistance to flow, a feature which may not be desirable in a particular study. Crack propagation studies may also be very sensitive to such linear correlations in the field. For applications such as these, the Turning Bands method should only be used with a sufficiently large number of lines. This may require some preliminary investigation for arbitrary covariance functions. In addition, the minimum number of lines in 3 and higher dimensions is difficult to determine due to visualization problems.

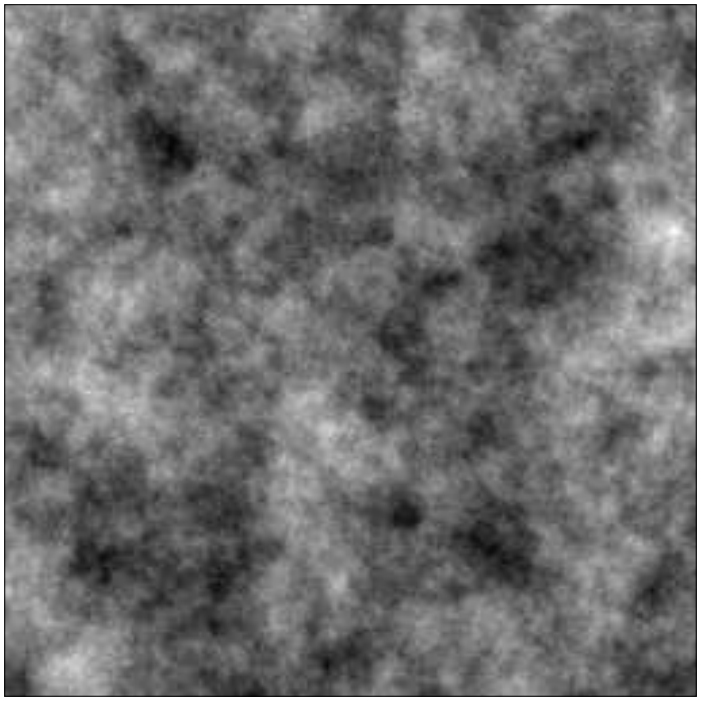


Figure 8. Sample function of a 2-D field via TBM using 64 lines.

Note that the Turning Bands Method does not suffer from the symmetric covariance structure that is inherent in the FFT approach. The variance field and covariance structure are also well preserved. However, the necessity of finding an equivalent 1-D covariance or spectral density function through an integral equation along with the streaked appearance of the realization when an insufficient number of lines are used makes the method less attractive. Using a larger number of lines, TBM is probably the most accurate of the three methods considered, at the expense of decreased efficiency, as long as the 1-D generator is accurate. TBM can be extended to anisotropic fields, although there is an additional efficiency penalty associated with such an extension since the 1-D process statistics must be recalculated for each new line orientation (see Mantaglou and Wilson, 1981, for details).

4.6 The Local Average Subdivision Method

Of the three approximate methods considered, the Local Average Subdivision (LAS) method is probably the most difficult to implement, but the easiest to use. LAS is a fast and generally accurate method of producing realizations of a discrete 'local average' random process. The motivation for the method arose out of a need to properly account for the fact that most engineering measurements are actually local averages of the property in question. For example, soil porosity is ill-defined at the micro-scale –

it is measured in practice using samples of finite volume and the measured value is an average of the porosity through the sample. The same can be said of strength measurements, say triaxial tests on laboratory volumes, or CPT measurements which record the effects of deforming a bulb of soil around the cone. The variance of the average is strongly affected by the size of the sample. Depending on the distribution of the property being measured, the mean of the average may also be affected by the sample size – this is sometimes called the *scale effect*. These effects are relatively easily incorporated into a properly defined random local average process.

Another advantage to using local averages is that they are ideally suited to stochastic finite element modeling using efficient, low order, interpolation functions. Each discrete local average given by a realization becomes the average property within each discrete element. As the element size is changed, the statistics of the random property mapped to the element will also change in a statistically consistent fashion. This gives finite element modelers the freedom to change mesh resolution without losing stochastic accuracy.

The concept behind the LAS approach derived from the stochastic subdivision algorithm described by Carpenter (1980) and Fournier et al. (1982). Their method was limited to modeling power spectra having a $\omega^{-\beta}$ form and suffered from problems with aliasing and 'creasing'. Lewis (1987) generalized the approach to allow the modeling of arbitrary power spectra without eliminating the aliasing. The stochastic subdivision is a midpoint displacement algorithm involve recursively subdividing the domain by generating new midpoint values randomly selected according to some distribution. Once chosen, the value at a point remains fixed and at each stage in the subdivision only half the points in the process are determined (the others created in previous iterations). Aliasing arises because the power spectral density is not modified at each stage to reflect the increasing Nyquist frequency associated with each increase in resolution. Voss (in Peitgen et al., 1988, Chap. 1) attempted to eliminate this problem with considerable success by adding randomness to all points at each stage in the subdivision in a method called 'successive random additions'. However the internal consistency easily achieved by the midpoint displacement methods (their ability to return to previous states while decreasing resolution through decimation) is largely lost with the successive random additions technique. The property of internal consistency in the midpoint displacement approaches implies that certain points retain their value throughout the subdivision and other points are created to remain consistent with them with respect to correlation. In the LAS approach, internal consistency implies that the local average is maintained throughout the subdivision.

The LAS method solves the problems associated with the stochastic subdivision methods and incorporates into it concepts of local averaging theory. The general concept and procedure is presented first for a one-dimensional stationary process characterized by its second-order statistics. The algorithm is illustrated by a Markov process, having a simple exponential correlation function, as well as by a fractional Gaussian noise process as defined by Mandelbrot and van Ness (1968). The simulation procedure in two and three dimensions is then described. Finally some comments concerning the

accuracy and efficiency of the method are made.

4.6.1 One-Dimensional Local Average Subdivision

The construction of a local average process via LAS essentially proceeds in a top-down recursive fashion as illustrated in Figure 9. In Stage 0, a global average is generated for the process. In Stage 1, the domain is subdivided into two regions whose 'local' averages must in turn average to the global (or parent) value. Subsequent stages are obtained by subdividing each 'parent' cell and generating values for the resulting two regions while preserving upwards averaging. Note that the global average remains constant throughout the subdivision, a property that is ensured merely by requiring that the average of each pair generated is equivalent to the parent cell value. This is also a property of any cell being subdivided. We note that the local average subdivision can be applied to any existing local average field. For example, the stage 0 shown in Figure 9 might simply be one local average cell in a much larger field. The algorithm proceeds as follows;

1. generate a normally distributed global average (labeled Z_1^0 in Figure 9) with mean zero and variance obtained from local averaging theory,
2. subdivide the field into two equal parts,
3. generate two normally distributed values, Z_1^1 and Z_2^1, whose means and variances are selected so as to satisfy three criteria:
   a) that they show the correct variance according to local averaging theory,
   b) that they are properly correlated with one another,
   c) that they average to the parent value, $\frac{1}{2}(Z_1^1 + Z_2^1) = Z_1^0$.
   That is, the distributions of Z_1^1 and Z_2^1 are conditioned on the value of Z_1^0,
4. subdivide each cell in stage 1 into two equal parts,
5. generate two normally distributed values, Z_1^2 and Z_2^2, whose means and variances are selected so as to satisfy four criteria:
   a) that they show the correct variance according to local averaging theory,
   b) that they are properly correlated with one another,
   c) that they average to the parent value, $\frac{1}{2}(Z_1^2 + Z_2^2) = Z_1^1$,
   d) that they are properly correlated with Z_3^2 and Z_4^2.
   The third criteria implies conditioning of the distributions of Z_1^2 and Z_2^2 on the value of Z_1^1. The fourth criteria will only be satisfied approximately by conditioning their distributions also on Z_2^1.

and so on in this fashion. The approximations in the algorithm come about in two ways: first the correlation with adjacent cells across parent boundaries is accomplished through the parent values (which are already known having been previously generated). Second the range of parent cells on which to condition the distributions will

be limited to some neighborhood. Much of the remainder of this section is devoted to the determination of these conditional Gaussian distributions at each stage in the subdivision and to an estimation of the algorithmic errors. In the following, the term 'parent cell' refers to the previous stage cell being subdivided and 'within-cell' means within the region defined by the parent cell.

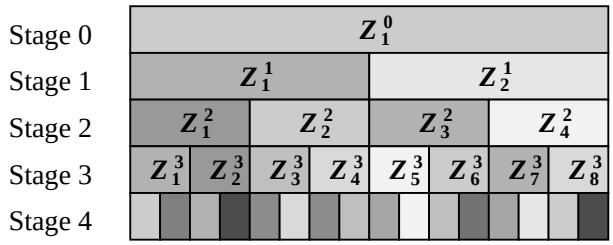


Figure 9. Top-down approach to LAS construction of local average random process.

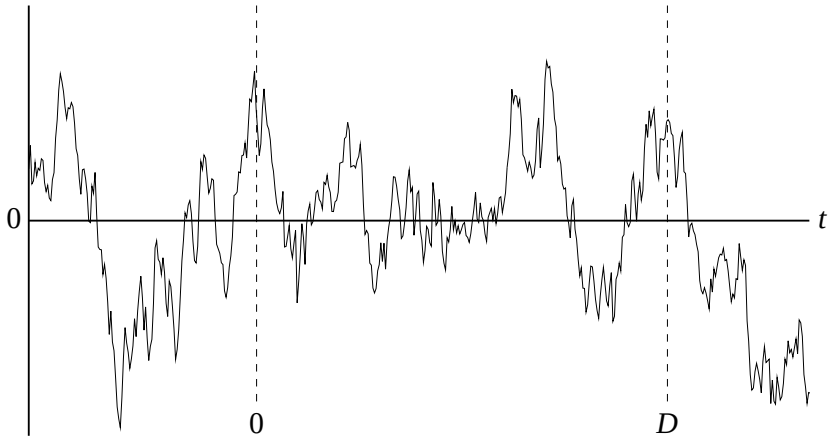


Figure 10. Realization of continuous random function, Z, with domain of interest $(0, D]$ shown.

To determine the mean and variance of the Stage 0 value, Z_1^0, consider first a continuous stationary scalar random function $Z(t)$ in one dimension, a sample of which may appear as shown in Figure 10, and define a domain of interest $(0, D]$ within which a realization is to be produced. Two comments should be made at this point: First, as it is currently implemented the LAS method is restricted to stationary processes fully described by their second-order statistics (mean, variance and correlation function or, equivalently, spectral density function). This is not a severe restriction since it leaves a sufficiently broad class of functions to model most natural phenomena (Lewis, 1987); also, there is often insufficient data to substantiate more complex probabilistic models. Besides, a non-stationary mean and variance can be easily added to a stationary process. For example $Y(t) = \mu(t) + \sigma(t) \times X(t)$ will produce a non-stationary $Y(t)$

from stationary $X(t)$ if $\mu(t)$ and/or $\sigma(t)$ vary with t. Secondly, the subdivision procedure depends on the physical size of the domain being defined since the dimension over which local averaging is to be performed must be known. The process Z beyond the domain $(0, D]$ is ignored.

The average of $Z(t)$ over the domain $(0, D]$ is given by

$$Z_1^0 = \frac{1}{D}\int_0^D Z(\xi)\, d\xi \tag{67}$$

where Z_1^0 is a random variable whose statistics

$$\mathrm{E}\big[Z_1^0\big] = \mathrm{E}\big[Z\big] \tag{68}$$

$$\begin{aligned}\mathrm{E}\big[(Z_1^0)^2\big] &= \frac{1}{D^2}\int_0^D \int_0^D \mathrm{E}\big[Z(\xi)\, Z(\xi')\big]\; d\xi\, d\xi' \\ &= \mathrm{E}\big[Z\big]^2 \;+\; \frac{2}{D^2}\int_0^D (D-\tau)\, C(\tau) d\tau\end{aligned} \tag{69}$$

can be found by making use of stationarity and the fact that $C(\tau)$, the covariance function of $Z(t)$, is an even function of lag τ. Without loss in generality, $\mathrm{E}\big[Z\big]$ will henceforth be taken as zero. If $Z(t)$ is a Gaussian random function, Eq's (68) and (69) give sufficient information to generate a realization of Z_1^0 which becomes stage 0 in the LAS method. If $Z(t)$ is not Gaussian, then the complete probability distribution function for Z_1^0 must be determined and a realization generated according to such a distribution. We will restrict our attention to Gaussian processes.

Consider now the general case where stage i is known and stage $i+1$ is to be generated. In the following the superscript i denotes the stage under consideration. Define

$$D^i = \frac{D}{2^i}, \qquad i = 0, 1, 2, \ldots, L, \tag{70}$$

where the desired number of subintervals in the final realization is $N = 2^L$, and define Z_k^i to be the average of $Z(t)$ over the interval $(k-1)D^i < t \le kD^i$ centered at $t_k = (k - \frac{1}{2})D^i$, i.e.

$$Z_k^i = \frac{1}{D^i}\int_{(k-1)D^i}^{kD^i} Z(\xi)\; d\xi \tag{71}$$

where $\mathrm{E}\big[Z_k^i\big] = \mathrm{E}\big[Z\big] = 0$. The target covariance between local averages separated by lag mD^i between centers is

$$\begin{aligned}\mathrm{E}\big[Z_k^i Z_{k+m}^i\big] &= \mathrm{E}\left[\left(\frac{1}{D^i}\right)^2 \int_{(k-1)D^i}^{kD^i} \int_{(k+m-1)D^i}^{(k+m)D^i} Z(\xi)\, Z(\xi')\; \mathrm{d}\xi \mathrm{d}\xi'\right] \\ &= \left(\frac{1}{D^i}\right)^2 \int_0^{D^i} \int_{mD^i}^{(m+1)D^i} B(\xi - \xi')\; \mathrm{d}\xi \mathrm{d}\xi'\end{aligned}$$

$$= \left(\frac{1}{D^i}\right)^2 \int_{(m-1)D^i}^{mD^i} \big(\xi - (m-1)D^i\big) B(\xi)\, \mathrm{d}\xi$$
$$+ \left(\frac{1}{D^i}\right)^2 \int_{mD^i}^{(m+1)D^i} \big((m+1)D^i - \xi\big) B(\xi) \mathrm{d}\xi. \tag{72}$$

which can be evaluated relatively simply using Gaussian quadrature as

$$\mathrm{E}\big[Z_k^i Z_{k+m}^i\big] \simeq \frac{1}{4} \sum_{\nu=1}^{n_g} w_\nu \Big[(1+z_\nu))C(r_\nu) + (1-z_\nu)C(s_\nu)\Big] \tag{73}$$

where $r_\nu = D^i\Big(m - \frac{1}{2}(1-z_\nu)\Big)$, $s_\nu = D^i\Big(m + \frac{1}{2}(1+z_\nu)\Big)$ and the weights, w_ν, and positions z_ν can be found in Appendix A.4 for n_g Gauss points.

With reference to Figure 11, the construction of stage $i+1$ given stage i is obtained by estimating a mean for Z_{2j}^{i+1} and adding a zero mean discrete white noise $c^{i+1}\, U_j^{i+1}$ having variance $(c^{i+1})^2$

$$Z_{2j}^{i+1} = M_{2j}^{i+1} + c^{i+1}\, U_j^{i+1}. \tag{74}$$

The best linear estimate for the mean M_{2j}^{i+1} can be determined by a linear combination of stage i (parent) values in some neighborhood $j-n, \ldots, j+n$,

$$M_{2j}^{i+1} = \sum_{k=j-n}^{j+n} a_{k-j}^i\, Z_k^i. \tag{75}$$

Multiplying (74) through by Z_m^i, taking expectations and using the fact that U_j^{i+1} is uncorrelated with the stage i values allows the determination of the coefficients a in terms of the desired covariances,

$$\mathrm{E}\big[Z_{2j}^{i+1}\, Z_m^i\big] = \sum_{k=j-n}^{j+n} a_{k-j}^i\, \mathrm{E}\big[Z_k^i\, Z_m^i\big] \tag{76}$$

a system of equations ($m = j-n, \ldots, j+n$) from which the coefficients a_ℓ^i, $\ell = -n, \ldots, n$, can be solved. The covariance matrix multiplying the vector $\{a_\ell^i\}$ is both symmetric and Toeplitz (elements along each diagonal are equal). For $U_j^{i+1} \sim N(0,1)$ the variance of the noise term is $(c^{i+1})^2$ which can be obtained by squaring (74), taking expectations and employing the results of (76)

$$(c^{i+1})^2 = \mathrm{E}\big[(Z_{2j}^{i+1})^2\big] - \sum_{k=j-n}^{j+n} a_{k-j}^i\, \mathrm{E}\big[Z_{2j}^{i+1}\, Z_k^i\big]\,. \tag{77}$$

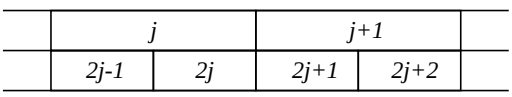


Figure 11. One-dimensional LAS indexing for stage i (top) and stage $i+1$ (bottom).

The adjacent cell, Z_{2j-1}^{i+1}, is determined by ensuring that upwards averaging is preserved – that the average of each stage $i+1$ pair equals the value of the stage i parent,

$$Z_{2j-1}^{i+1} = 2\, Z_j^i - Z_{2j}^{i+1} \tag{78}$$

which incidentally gives a means of evaluating the cross-stage covariances

$$\mathrm{E}\big[Z_{2j}^{i+1}\, Z_m^i\big] = \tfrac{1}{2}\mathrm{E}\big[Z_{2j}^{i+1}\, Z_{2m-1}^{i+1}\big] + \tfrac{1}{2}\mathrm{E}\big[Z_{2j}^{i+1}\, Z_{2m}^{i+1}\big]\,. \tag{79}$$

which are needed in Eq. 76. All the expectations in Equations (76) to (79) are evaluated using (72) or (73) at the appropriate stage.

For stationary processes, the set of coefficients $\{a_\ell^i\}$ and c^i are independent of position since the expectations in (76) and (77) are just dependent on lags. The generation procedure can be restated as follows;

1. for $i = 0, 1, 2, \ldots, L$ compute the coefficients $\{a_\ell^i\},\ \ell = -n, \ldots, n$ using (76) and c^{i+1} using (77),
2. starting with $i = 0$, generate a realization for the global mean using (68) and (69),
3. subdivide the domain,
4. for each $j = 1, 2, 3, \ldots, 2^i$, generate realizations for Z_{2j}^{i+1} and Z_{2j-1}^{i+1} using (74) and (78),
5. increment i and, if not greater than L, return to step 3.

Notice that subsequent realizations of the process need only start at step 2 and so the overhead involved with setting up the coefficients becomes rapidly negligible.

Because the LAS procedure is recursive, obtaining stage $i+1$ values using the previous stage, it is relatively easy to condition the field by specifying the values of the local averages at a particular stage. So, for example, if the global mean of a process is known *a priori*, then the stage 0 value can be set to this mean and the LAS procedure started at stage 1. Similarly if the resolution is to be refined in a certain region, then the values in that region become the starting values and the subdivision resumed at the next stage.

Although the LAS method yields a local average process, when the discretization size becomes small enough it is virtually indistinguishable from the limiting continuous process. Thus the method can be used to approximate continuous functions as well.

Accuracy

It is instructive to investigate how closely the algorithm approximates the target statistics of the process. Changing notation slightly, denote the stage $i+1$ algorithmic values, given the stage i values, as

$$\hat{Z}_{2j}^{i+1} = c^{i+1}\, U_j^{i+1} + \sum_{k=j-n}^{j+n} a_{k-j}^i\, Z_k^i \tag{80}$$

$$\hat{Z}_{2j-1}^{i+1} = 2\, Z_j^i - \hat{Z}_{2j}^{i+1}. \tag{81}$$

It is easy to see that the expectation of $\hat{Z}$ is still zero, as desired, while the variance is

$$\begin{aligned}
\mathrm{E}\Big[(\hat{Z}_{2j}^{i+1})^2\Big] &= \mathrm{E}\left[\left(c^{i+1}\, U_j^{i+1} + \sum_{k=j-n}^{j+n} a_{k-j}^i\, Z_k^i\right)^2\right] \\
&= (c^{i+1})^2 + \sum_{k=j-n}^{j+n} a_{k-j}^i \sum_{\ell=j-n}^{j+n} a_{\ell-j}^i\, \mathrm{E}\big[Z_k^i\, Z_\ell^i\big] \\
&= \mathrm{E}\big[(Z_{2j}^{i+1})^2\big] - \sum_{k=j-n}^{j+n} a_{k-j}^i\, \mathrm{E}\big[Z_{2j}^{i+1}\, Z_k^i\big] + \sum_{k=j-n}^{j+n} a_{k-j}^i\, \mathrm{E}\big[Z_{2j}^{i+1}\, Z_k^i\big] \\
&= \mathrm{E}\big[(Z_{2j}^{i+1})^2\big]
\end{aligned} \tag{82}$$

in which the coefficients c^{i+1} and a_ℓ^i where calculated using (76) and (77) as before. Similarly, the within-cell covariance at lag D^{i+1} is

$$\begin{aligned}
\mathrm{E}\Big[\hat{Z}_{2j-1}^{i+1}\, \hat{Z}_{2j}^{i+1}\Big] &= 2 \sum_{\ell=j-n}^{j+n} a_{\ell-j}^i\, \mathrm{E}\big[Z_\ell^i\, Z_j^i\big] - \mathrm{E}\big[(Z_{2j}^{i+1})^2\big] \\
&= 2\, \mathrm{E}\big[Z_{2j}^{i+1}\, Z_j^i\big] - \mathrm{E}\big[(Z_{2j}^{i+1})^2\big] \;=\; \mathrm{E}\big[Z_{2j-1}^{i+1}\, Z_{2j}^{i+1}\big]
\end{aligned} \tag{83}$$

using the results of (82) along with (79). Thus the covariance structure within a cell is preserved *exactly* by the subdivision algorithm. Some approximation does occur across cell boundaries as can be seen by considering

$$\begin{aligned}
\mathrm{E}\Big[\hat{Z}_{2j}^{i+1}\, \hat{Z}_{2j+1}^{i+1}\Big] &= 2 \sum_{k=j-n}^{j+n} a_{k-j}^i\, \mathrm{E}\big[Z_k^i\, Z_{j+1}^i\big] \\
&\quad - \sum_{\ell=j-n+1}^{j+n+1} a_{\ell-j-1}^i \sum_{k=j-n}^{j+n} a_{k-j}^i\, \mathrm{E}\big[Z_k^i\, Z_\ell^i\big] \\
&= \mathrm{E}\big[Z_{2j}^{i+1}\, Z_{2j+1}^{i+1}\big] + \mathrm{E}\big[Z_{2j}^{i+1}\, Z_{2j+2}^{i+1}\big] \\
&\quad - \sum_{\ell=j-n+1}^{j+n+1} a_{\ell-j-1}^i \mathrm{E}\big[Z_{2j}^{i+1}\, Z_\ell^i\big]
\end{aligned} \tag{84}$$

The algorithmic error in this covariance comes from the last two terms. The discrepancy between (84) and the exact covariance is illustrated numerically in Figure 12 for a zero mean Markov process having covariance and variance functions

$$B(\tau) = \sigma^2\, \exp\left\{-\frac{2|\tau|}{\theta}\right\} \tag{85}$$

$$\gamma(T) = \frac{\theta^2}{2\,T^2}\left[\frac{2|T|}{\theta} + \exp\left\{\frac{-2|T|}{\theta}\right\} - 1\right] \tag{86}$$

where T is the averaging dimension (in Figure 12, $T = D^{i+1}$) and θ is the scale of fluctuation of the process. The exact covariance is determined by (72) (for $m = 1$) using the variance function (86). Although Figure 12 shows a wide range in the effective cell sizes, $2T/\theta$, the error is typically very small.

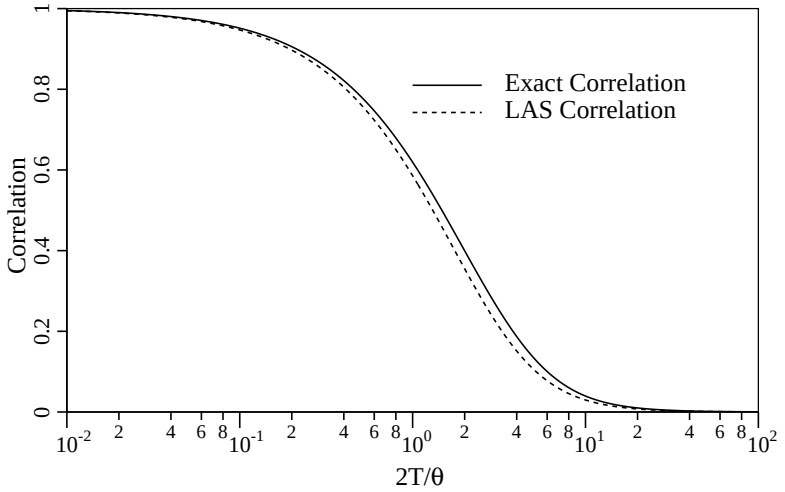


Figure 12. Comparison of algorithmic and exact correlation between adjacent cells across a parent cell boundary for varying effective cell dimension $2T/\theta$.

To address the issue of errors at larger lags and the possibility of errors accumulating from stage to stage, it is useful to look at the exact versus estimated statistics of the entire process. Figure 13 illustrates this comparison for the Markov process. It can be seen from this example and from the fractional Gaussian noise example to come, that the errors are self-correcting and the algorithmic correlation structure tends to the exact correlation function when averaged over several realizations. Spectral analysis of realizations obtained from the LAS method show equally good agreement between estimated and exact (Fenton, trefFent90). The within-cell rate of convergence of the estimated statistics to the exact is $1/n_{sim}$, where n_{sim} is the number of realizations. The overall rate of convergence is about the same.

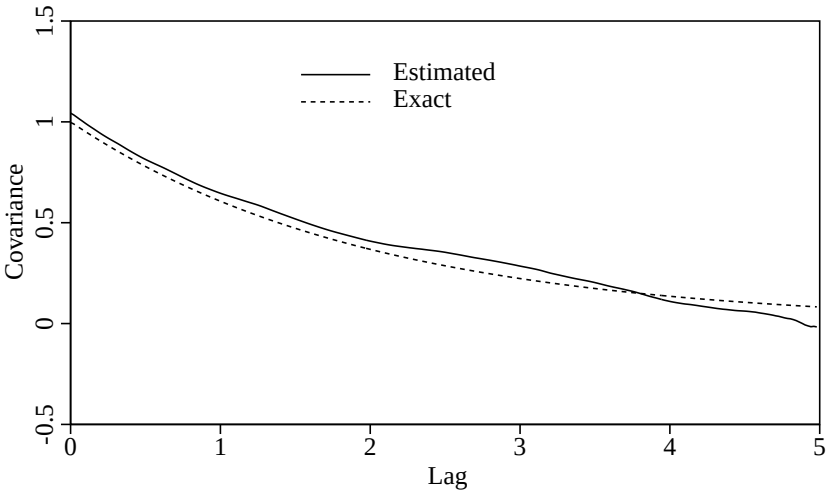


Figure 13. Comparison of exact and estimated covariance functions (averaged over 200 realizations) of a Markov process with $\sigma = 1$ and $\theta = 4$.

Boundary Conditions and Neighborhood Size

When the neighborhood size $(2n + 1)$ is greater than 1 $(n > 0)$, the construction of values near the boundary may require values from the previous stage which lie outside the boundary. This problem is handled by assuming that what happens outside the domain $(0, D]$ is of no interest and uncorrelated with what happens within the domain. The generating relationship (74) near either boundary becomes

$$Z_{2j}^{i+1} = c^{i+1}\, U_j^{i+1} + \sum_{k=j-p}^{j+q} a_{k-j}^i\, Z_k^i \tag{87}$$

where $p = \min(n, j-1)$, $q = \min(n, 2^i - j)$ and the coefficients a_ℓ^i need only be determined for $\ell = -p, \ldots, q$. The periodic boundary conditions mentioned by Lewis (1987) are not appropriate if the target covariance structure is to be preserved since they lead to a covariance which is symmetric about lag $D/2$ (unless the desired covariance is also symmetric about this lag).

In the implementation described in this paper, a neighborhood size of 3 was used $(n = 1)$, the parent cell plus its two adjacent cells. Because of the top-down approach, there seems to be little justification to using a larger neighborhood for processes with covariance functions which decrease monotonically or which are relatively smooth. When the covariance function is oscillatory, a larger neighborhood is required in order to successfully approximate the function.

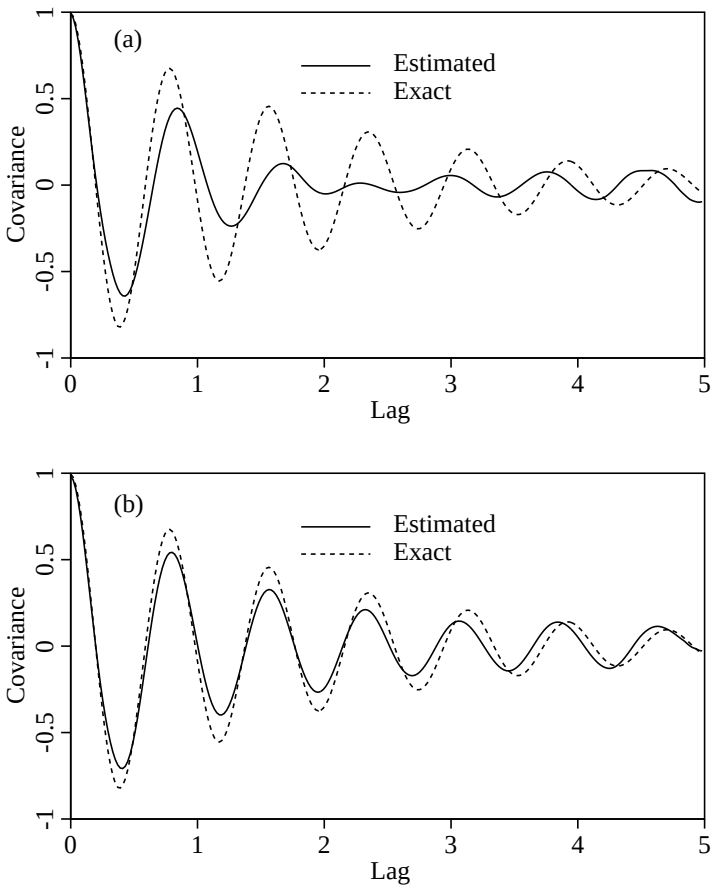


Figure 14. Effect of neighborhood size for (a) $n = 1$ and (b) $n = 2$ for a damped oscillatory process.

In Figure 14 the exact and estimated covariances are shown for a damped oscillatory process where

$$B(\tau) = \sigma^2 \cos(\omega\tau)\, \mathrm{e}^{-2\tau/\theta}. \tag{88}$$

Considerable improvement in the model is obtained when a neighborhood size of 5 is used $(n = 2)$. This improvement comes at the expense of taking about twice as long to generate the realizations. Many practical models of natural phenomena employ monotonically decreasing covariance functions, often for simplicity, and so the $n = 1$ implementation is usually preferable.

Fractional Gaussian Noise

As a further demonstration of the LAS method, a self-similar process called fractional Gaussian noise was simulated. Fractional Gaussian noise (fGn) is defined by Mandelbrot et al. (1968) to be the derivative of fractional Brownian motion (fBm), and is obtained by averaging the fBm over a small interval δ. The resulting process has

covariance and variance functions

$$B(\tau) = \frac{\sigma^2}{2\delta^{2H}}\left[|\tau+\delta|^{2H} - 2|\tau|^{2H} + |\tau-\delta|^{2H}\right] \tag{89}$$

$$\gamma(T) = \frac{|T+\delta|^{2H+2} - 2|T|^{2H+2} + |T-\delta|^{2H+2} - 2\delta^{2H+2}}{T^2(2H+1)(2H+2)\delta^{2H}} \tag{90}$$

defined for $0 < H < 1$.

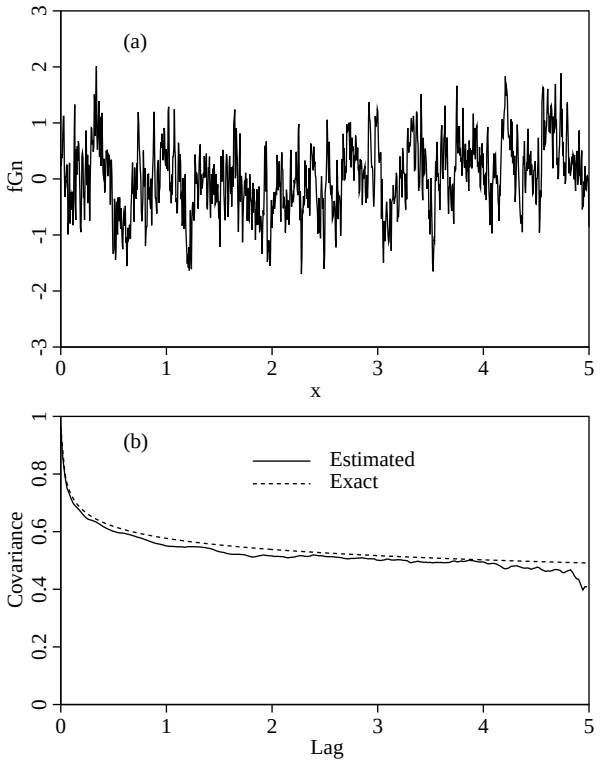


Figure 15. (a) LAS generated sample function of ω^{2H-1} noise for $H = 0.95$, and (b) corresponding estimated (averaged over 200 realizations) and exact covariance functions.

The case $H = 0.5$ corresponds to white noise and $H \to 1$ gives ω^{-1} type noise. In practice δ is taken to be equal to the smallest lag between field points ($\delta = D/2^L$) to ensure that when $H = 0.5$ (white noise), $B(\tau)$ becomes zero for all $\tau \geq D/2^L$. A sample function and its corresponding ensemble statistics are shown in Figure 15 for $\omega^{-\beta}$ type noise ($H = 0.95$) where $\beta = 2H - 1$. The self-similar type processes have been demonstrated by Mandelbrot (1982), Voss (1985), and many others (Mohr, 1981, Peitgen et al., 1988, Whittle, 1956, to name a few) to be representative of a large variety of natural forms and patterns, for example music, terrains, crop yields, and chaotic systems. Fenton (1999) demonstrated the presence of fractal behaviour in CPT logs taken in Norway.

4.6.2 Multi-Dimensional Local Average Subdivision

The 2-D LAS method involves a subdivision process in which a 'parent' cell is divided into 4 equal sized cells. In Figure 16, the parent cells are denoted Z_l^i, $l = 1, 2, \ldots$ and the subdivided, or child cells are denoted Z_j^{i+1}, $j = 1, 2, 3, 4$. Although each parent cell is eventually subdivided in the LAS process, only Z_5^i is subdivided in Figure 16 for simplicity. Using vector notation, the values of the column vector $\underset{\sim}{Z}^{i+1} = \{Z_1^{i+1}, Z_2^{i+1}, Z_3^{i+1}, Z_4^{i+1}\}$ are obtained by adding a mean term to a random component. The mean term derives from a best linear unbiased estimate using a 3×3 neighborhood of the parent values, in this case the column vector $\underset{\sim}{Z}^i = \{Z_1^i, \ldots, Z_9^i\}$. Specifically

$$\underset{\sim}{Z}^{i+1} = \underset{\approx}{A}^T \underset{\sim}{Z}^i + \underset{\approx}{L}\underset{\sim}{U} \tag{91}$$

where $\underset{\sim}{U}$ is a random vector with independent $N(0, 1)$ elements. This is essentially an ARMA model in which the 'past' is represented by the previous coarser resolution stages.

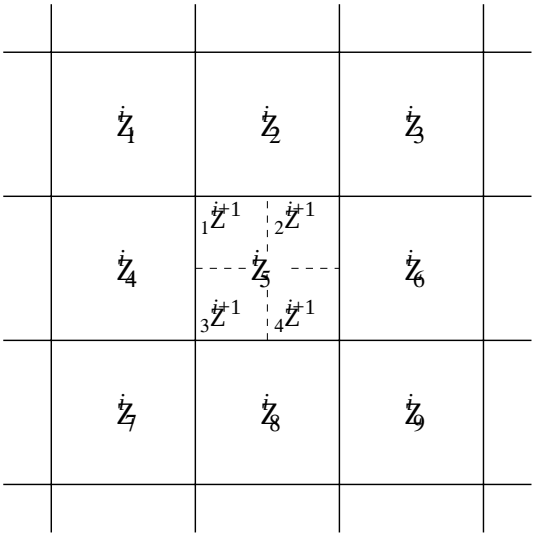


Figure 16. Local Average Subdivision in two-dimensions.

Defining the covariance matrices

$$\underset{\approx}{R} = \mathrm{E}\left[\underset{\sim}{Z}^i \underset{\sim}{Z}^{i^T}\right], \tag{92a}$$

$$\underset{\approx}{S} = \mathrm{E}\left[\underset{\sim}{Z}^i \underset{\sim}{Z}^{i+1^T}\right], \quad \text{and} \tag{92b}$$

$$\underset{\approx}{B} = \mathrm{E}\left[\underset{\sim}{Z}^{i+1} \underset{\sim}{Z}^{i+1^T}\right], \tag{92c}$$

then the matrix $\underset{\approx}{A}$ is determined by

$$\underset{\approx}{A} = \underset{\approx}{R}^{-1} \underset{\approx}{S} \tag{93}$$

while the lower triangular matrix $\underset{\approx}{L}$ satisfies

$$\underset{\approx}{L}\underset{\approx}{L}^T = \underset{\approx}{B} - \underset{\approx}{S}^T\underset{\approx}{A} \tag{94}$$

The covariance matrices $\underset{\approx}{R}$, $\underset{\approx}{S}$ and $\underset{\approx}{B}$ must be computed as the covariances between local averages over the domains of the parent and child cells. This can be done using the variance function although direct Gauss Quadrature integration of the covariance function has been found to give better numerical results.

Note that the matrix on the right hand side of Eq. 94 is only rank 3, so that the 4×4 matrix $\underset{\approx}{L}$ has a special form with columns summing to zero (thus $L_{44} = 0$). While this results from the fact that all the expectations used in Eq.'s 92 are derived using local average theory over the cell domains, the physical interpretation is that upwards averaging is preserved, ie. that $P_5 = \frac{1}{4}(Q_1 + Q_2 + Q_3 + Q_4)$. This means that one of the elements of $\underset{\sim}{Q}$ is explicitly determined once the other three are known. In detail, Eq. 91 is carried out as follows

$$Z_1^{i+1} = \sum_{l=1}^{9} A_{l1} Z_l^i + L_{11}U_1 \tag{95a}$$

$$Z_2^{i+1} = \sum_{l=1}^{9} A_{l2} Z_l^i + L_{21}U_1 + L_{22}U_2 \tag{95b}$$

$$Z_3^{i+1} = \sum_{l=1}^{9} A_{l3} Z_l^i + L_{31}U_1 + L_{32}U_2 + L_{33}U_3 \tag{95c}$$

$$Z_4^{i+1} = 4Z_5^i - Z_1^{i+1} - Z_2^{i+1} - Z_3^{i+1} \tag{95d}$$

where U_i are a set of three independent standard normally distributed random variables. Subdivisions taking place near the field boundaries are handled in much the same manner as in the one-dimensional case by assuming that conditions outside the field are uncorrelated with those inside the field.

The assumption of homogeneity vastly decreases the number of coefficients that need to be calculated and stored since the matrices $\underset{\approx}{A}$ and $\underset{\approx}{L}$ become independent of position. As in the 1-D case, the coefficients need only be calculated prior to the first realization – they can be re-used in subsequent realizations reducing the effective cost of their calculation.

A sample function of a Markov process having isotropic covariance function

$$B(\tau_1, \tau_2) = \sigma^2 \exp\{-\tfrac{2}{\theta}\sqrt{\tau_1^2 + \tau_2^2}\} \tag{96}$$

was generated using the two-dimensional LAS algorithm and is shown in Figure 17. The field, which is of dimension 5×5, was subdivided 8 times to obtain a 256×256 resolution giving relatively small cells of size $\frac{5}{256} \times \frac{5}{256}$. The estimated covariances along three different directions are seen in Figure 18 to show very good agreement

with the exact. The agreement improves (as $1/n_{sim}$) when the statistics are averaged over a larger number of simulations. Notice that the horizontal axis on Figure 18 extends beyond a lag of 5 to accomodate the estimation of the covariance along the diagonal (which has length $5\sqrt{2}$).

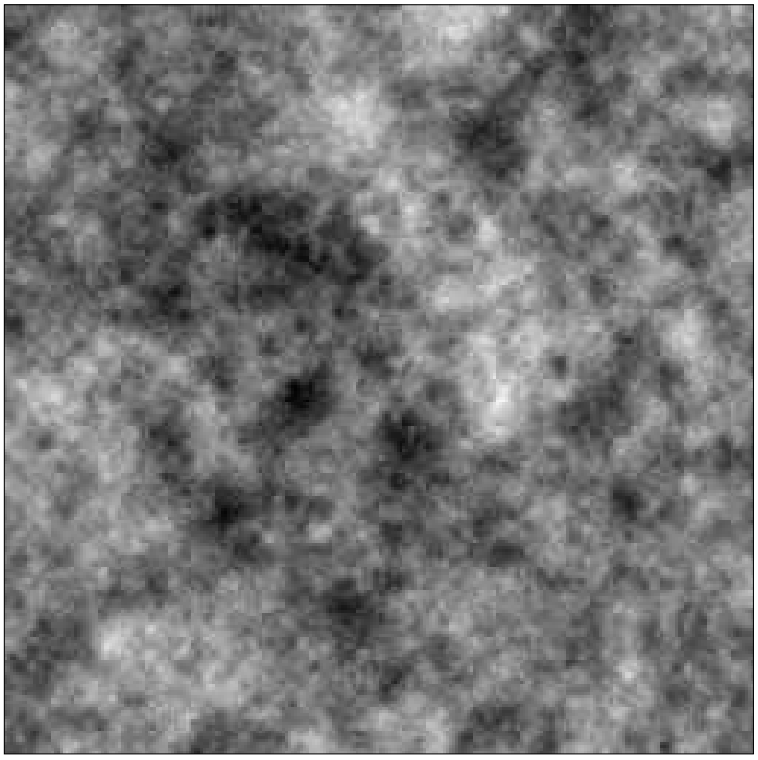


Figure 17. Local Average Subdivision generated two-dimensional sample function with $\theta = 0.5$.

In three dimensions, the LAS method involves recursively subdividing rectangular parallelepipeds into 8 equal volumes at each stage. The generating relationships are essentially the same as in the 2-D case except now 7 random noises are used in the subdivision of each parent volume at each stage

$$Z_s^{i+1} = \sum_{l=1}^{27} A_{ls} Z_l^i + \sum_{r=1}^{s} L_{sr}\, U_r \qquad s = 1, 2, \ldots, 7 \tag{97}$$

$$Z_8^{i+1} = 8 Z_{14}^i - \sum_{s=1}^{7} Z_s^{i+1} \tag{98}$$

in which Z_s^{i+1} denotes a particular octant of the subdivided cell centered at Z_{14}^i. Eq. 97 assumes a neighborhood size of $3 \times 3 \times 3$, and the subdivided cell is Z_{14}^i at the center of the neighborhood.

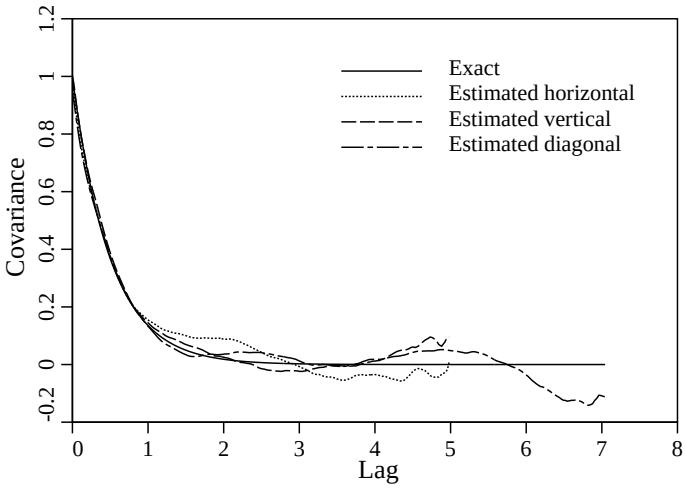


Figure 18. Comparison of exact and estimated covariance functions (averaged over 100 realizations) of a two-dimensional isotropic Markov process with $\sigma = 1$ and $\theta = 0.5$.

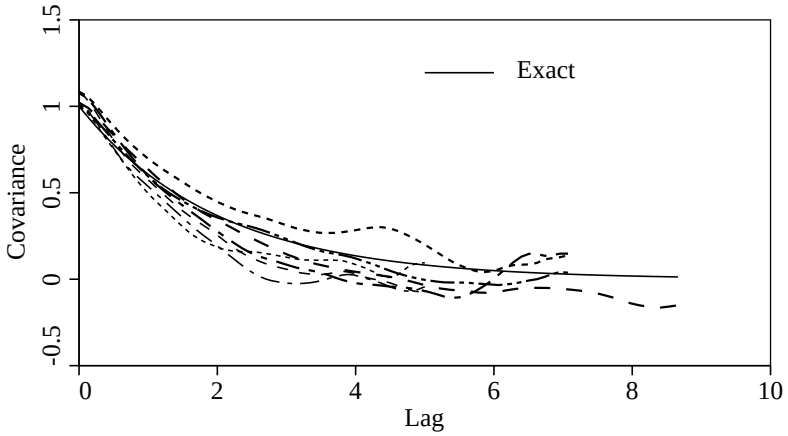


Figure 19. Comparison of exact and estimated covariance functions (averaged over 50 realizations) of a three-dimensional isotropic Markov process with $\sigma = 1$ and $\theta = 0.5$. The dashed lines show covariance estimates in various directions through the field.

Figure 19 compares the estimated and exact covariance of a three-dimensional first-order Markov process having isotropic covariance

$$B(\tau_1, \tau_2, \tau_3) = \sigma^2 \exp\{-\tfrac{2}{\theta}\sqrt{\tau_1^2 + \tau_2^2 + \tau_3^2}\} \tag{99}$$

The physical field size of $5 \times 5 \times 5$ was subdivided 6 times to obtain a resolution of $64 \times 64 \times 64$ and the covariance estimates were averaged over 50 realizations.

4.6.3 Implementation and Accuracy

In order to calculate stage $i+1$ values, the values at stage i must be known. This implies that in the 1-D case, storage must be provided for at least $1.5\,N$ values where $N = 2^L$ is the desired number of intervals of the process. If rapid 'zooming out' of the field is desired, it is useful to store all previous stages. This results in a storage requirement of $(2N-1)$ in 1-D, $\frac{4}{3}(N \times N)$ in 2-D, and $\frac{8}{7}(N \times N \times N)$ in 3-D. The coefficients $\underset{\approx}{A}$ and the lower triangular elements of $\underset{\approx}{L}$, which must also be stored, can be efficiently calculated using Gaussian elimination and Cholesky decomposition, respectively.

In two and higher dimensions, the LAS method, as presented above with a neighborhood size of 3 in each direction, is incapable of preserving anisotropy in the covariance structure. The directional scales of fluctuation tend toward the minimum for the field. To overcome this problem, the LAS method can be mixed with the covariance matrix decomposition (CMD) method (see Eq. 28). As mentioned in Section 4.3, the CMD method requires large amounts of storage and is prone to numerical error when the field to be simulated is not small. However, the first several stages of the local average field could be produced directly by the CMD method and then refined by LAS in subsequent stages until the desired field resolution is obtained. The resulting field would have anisotropy preserved at the large scale.

Specifically, in the one dimensional case, a positive integer k_1 is found so that the total number of cells, N_1, desired in the final field can be expressed as

$$N_1 = k_1(2^m) \tag{100}$$

where m is the number of subdivisions to perform and k_1 is as large as possible with $k_1 \leq k_{max}$. The choice of the upper bound k_{max} depends on how large the initial covariance matrix used in Eq. 28 can be. If k_{max} is too large, the Cholesky decomposition of the initial covariance matrix will be prone to numerical errors and algorithmic non-positive definiteness (which means that the Cholesky decomposition will fail). The authors suggest $k_{max} \leq 256$.

In two dimensions, two positive integers k_1 and k_2 are found such that $k_1 k_2 \leq k_{max}$ and the field dimensions can be expressed as

$$N_1 = k_1(2^m) \tag{101a}$$
$$N_2 = k_2(2^m) \tag{101b}$$

from which the first $k_1 \times k_2$ lattice of cell values are simulated directly using covariance matrix decomposition (28). Since the number of subdivisions, m, is common to the two parameters, one is not entirely free to choose N_1 and N_2 arbitrarily. It does, however, give a reasonable amount of discretion in generating non-square fields, as is also possible with both the FFT and TBM methods.

Although Figure 4 illustrates the superior performance of the LAS method over the FFT method in one dimension with respect to the covariance, a systematic bias in the

variance field is observed in two dimensions. Figure 20 shows a grey scale image of the estimated cell variance in a two-dimensional field obtained by averaging over the ensemble. There is a pattern in the variance field – the variance tends to be lower near the major cell divisions, that is at the $1/2$, $1/4$, $1/8$, etc. points of the field. This is because the actual diagonal, or variance, terms of the 4×4 covariance matrix corresponding to a subdivided cell are affected by the truncation of the parent cell influence to a 3×3 neighborhood. The error in the variance is compounded at each subdivision stage and cells close to 'older' cell divisions show more error than do 'interior' cells. The magnitude of this error varies with the number of subdivisions, the scale of fluctuation, and type of covariance function governing the process.

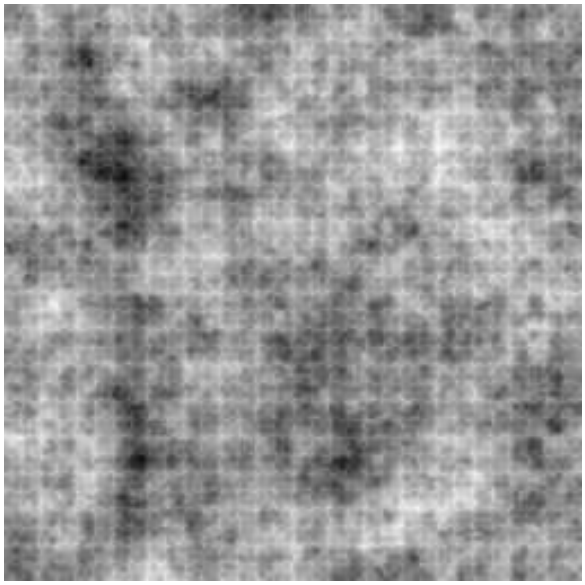


Figure 20. Two-dimensional LAS generated variance field (averaged over 200 realizations).

Figure 21 depicts the estimated variances along a line through the plane for both the LAS and TBM methods. Along any given line, the pattern in the LAS estimated variance seen in Figure 20 is not particularly noticeable and the values are about what would be expected for an estimate over the ensemble. Figure 22 compares the estimated covariance structure in the vertical and horizontal directions, again for the TBM (64 lines) and LAS methods. In this respect, both the LAS and the TBM methods are reasonably accurate.

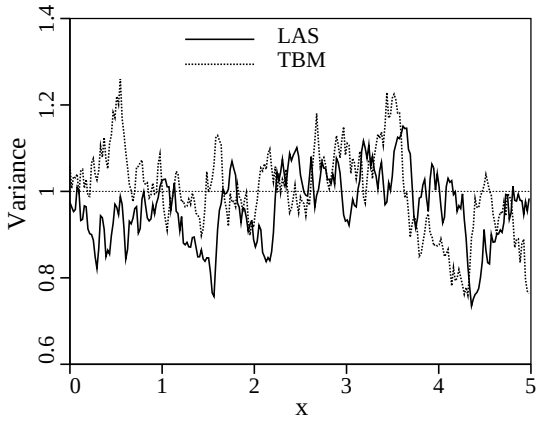


Figure 21. Variance along a horizontal line through the two-dimensional LAS and TBM fields estimated over 200 realizations.

Figure 23 illustrates how well the LAS method combined with the Covariance Matrix Decomposion (CMD) preserves anisotropy in the covariance structure. In this figure the horizontal scale of fluctuation is $\theta_x = 10$ while the vertical scale of fluctuation is $\theta_y = 1$. As mentioned earlier the LAS algorithm, using a neighborhood size of 3, is incapable of preserving anistropy. The anisotropy seen in Figure 23 is due to the initial CMD. The loss of anisotropy at very small lags (at the smaller scales where the subdivision is taking place) can be seen in the Figure – that is, the estimated horizontal covariance initially drops too rapidly at small lags.

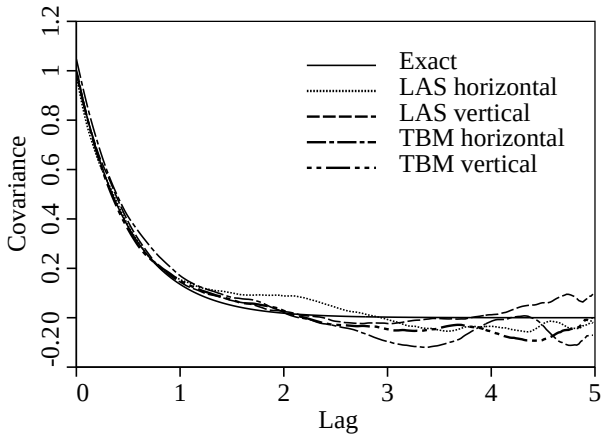


Figure 22. Covariance structure of the LAS and TBM two-dimensional random fields estimated over 200 realizations.

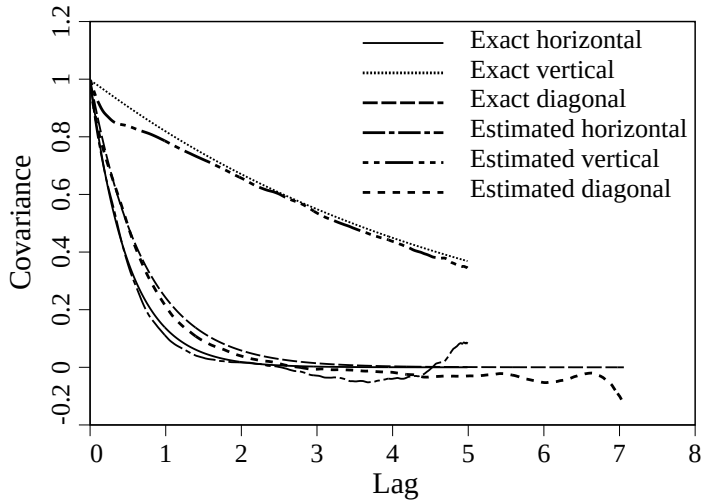


Figure 23. Exact and estimated covariance structure of an anisotropic LAS produced field with $\theta_x = 10$ and $\theta_y = 1$. The estimation is over 500 realizations.

It may be possible to improve the LAS covariance approximations by extending the size of the parent cell neighborhood. A 3×3 neighborhood is used in the current implementation of the 2-D LAS algorithm, as shown in Figure 16, but any odd sized neighborhood could be used to condition the statistics of the subdivided cells. Larger neighborhoods have not been tested in two and higher dimensions, although in one dimension increasing the neighborhood size to 5 cells resulted in a more accurate covariance function representation, as would be expected.

4.7 Comparison of Methods

The choice of a random field generator to be used for a particular problem or in general depends on many issues. Table 1 shows the relative run times of the three algorithms to produce identically sized fields. The times have been normalized with respect to the FFT method so that a value of 2 indicates that the method took twice as long as did the FFT. If efficiency alone were the selection criteria, then either the TBM with a small number of lines or the LAS methods would be selected, with probably the LAS a better choice if streaking is not desired. However, efficiency of the random field generator is often not an overriding concern – in many applications, the time taken to generate the field is dwarfed by the time taken to subsequently process or analyze the field. Substantial changes in generator efficiency may be hardly noticed by the user.

Table 1 Comparison of run-times of the FFT, TBM and LAS algorithms in one and two-dimensions.

Dimension	FFT	LAS	TBM	
			16 lines	64 lines
1-D	1.0	0.70	–	–
2-D	1.0	0.55	0.64	2.6

As a further comparison of the accuracy of the FFT, TBM, and LAS methods, a set of 200 realizations of a 128×128 random field were generated using the Markov covariance function with a scale of fluctuation $\theta = 2$ and a physical field size of 5×5. The mean and variance fields were calculated by estimating these quantities at each point in the field (averaging over the ensemble) for each algorithm. The upper and lower 90% quantiles are listed in Table 2 along with those predicted by theory under a normal distribution. To obtain these numbers, the mean and variance fields were first estimated, then upper and lower bounds were found such that 5% of the field exceeded the bounds above and below, respectively. Thus 90% of the field is observed to lie between the bounds. It can be seen that all three methods yield very good results with respect to the expected mean and variance quantiles. The TBM results were obtained using 64 lines. Although these results are strictly only valid for the particular covariance function used, they are believed to be generally true over a wider variety of covariance functions and scales of fluctuation.

Table 2 Upper and lower 90% quantiles of the estimated mean and variance fields for the FFT, TBM, and LAS methods (200 realizations).

Algorithm	Mean	Variance
FFT	$(-0.06, 0.12)$	$(0.87, 1.19)$
TBM	$(-0.11, 0.06)$	$(0.83, 1.14)$
LAS	$(-0.12, 0.09)$	$(0.82, 1.13)$
Theory	$(-0.12, 0.12)$	$(0.84, 1.17)$

Purely on the basis of accuracy in the mean, variance and covariance structures, the best algorithm of those considered here is probably the TBM method using a large number of lines. The TBM method is also one of the easiest to implement once an accurate 1-D generator has been implemented. Unfortunately, there is no clear rule regarding the minimum number of lines to be used to avoid streaking. In two dimensions using the Markov covariance function, it appears that at least 50 lines should be employed. However, as mentioned, narrow band processes may require more. In three dimensions, no such statements can be made due to the difficulty in studying the streaking phenomena off a plane. Presumably one could use a 'density' of lines similar to that used in the two-dimensional case, perhaps subtending similar angles, as a

guide. The TBM method is reasonably easy to use in practice as long as the equivalent 1-D covariance or spectral density function can be found.

The FFT method suffers from symmetry in the covariance structure of the realizations. This can be overcome by generating fields twice as large as required in each coordinate direction and ignoring the surplus. This correction results in slower run times (a factor of 2 in 1-D, 4 in 2-D, etc.). The FFT method is also relatively easy to implement and the algorithm is similar in any dimension. Its ability to easily handle anisotropic fields makes it the best choice for such problems. Care must be taken when selecting the physical field dimension and discretization interval to ensure that the spectral density function is adequately approximated. This latter issue makes the method more difficult to use in practice. However, the fact that the FFT approach employs the spectral density function directly makes it an intuitively attractive method, particularly in time dependent applications.

The LAS method has a systematic bias in the variance field, in two and higher dimensions, which is not solvable without increasing the parent neighborhood size. However, the error does not result in values of variance that lie outside what would be expected from theory – it is primarily the pattern of the variance field which is of concern. Of the three methods considered, the LAS method is the most difficult to implement. It is, however, one of the easiest to use once coded since it requires no decisions regarding its parameters, and it is generally the most efficient. If the problem at hand requires or would benefit from a local average representation, then the LAS method is the logical choice.

5 References

Box, G.E.P. and Muller, M.E. (1958). "A note on the generation of random normal variates," *Ann. Math. Statist.*, **29**, 610–611.

Carpenter, L. (1980). "Computer Rendering of fractal curves and surfaces," *SIGGRAPH 80 Proceedings*, ACM, New York, Seattle, Washington, 108–120.

Cooley, J.W. and Tukey, J.W. (1965). "An Algorithm for the Machine Calculation of Complex Fourier Series," *Mathematics of Computation*, **19**(90), 297-301.

Fenton, G.A. (1999). "Random field modeling of CPT data," *ASCE J. Geotech. Geoenv. Engrg.*, **125**(6), 486–498.

Fournier, A., Fussell, D. and Carpenter, L. (1982). "Computer rendering of stochastic models," *Commun. ACM*, **25**(6), 371-384.

Hull, T.E. and Dobell, A.R. (1962). "Random number generators," *SIAM Review*, **4**, 230–254.

Journel, A.G. and Huijbregts, Ch.J. (1978). *Mining Geostatistics*, Academic Press, New York, NY.

Knuth, D.E. (1981). *Seminumerical algorithms*, Vol. 2 of *The Art of Computer Programming*, (2nd Ed.), Addison-Wesley, Reading, MA.

L'Ecuyer, P. (1988). "Efficient and portable combined random number generators," *Communications of the ACM*, **31**, 742–749 and 774.

Law, A.M. and Kelton, W.D. (2000). *Simulation Modeling and Analysis*, (3rd Ed.), McGraw-Hill, New York, NY.

Lehmer, D.H. (1951). "Mathematical methods in large-scale computing units," *Ann. Comput. Lab.*, **26**, Harvard Univ., 141–146.

Lewis, J.P. (1987). "Generalized Stochastic Subdivision," *ACM Transactions on Graphics*, **6**(3), 167-190.

Mandelbrot, B.B. (1982). *The Fractal Geometry of Nature*, W.H. Freeman and Co., New York, NY.

Mandelbrot, B.B. and Ness, J.W. (1968). "Fractional Brownian Motions, Fractional Noises and Applications," *SIAM Review*, **10**(4), 422-437.

Mantoglou, A. and Wilson, J.L. (1981). "Simulation of Random Fields with the Turning Bands Method" , MIT, Dept. Civil Engrg., Report #264, Cambridge, MA.

Mohr, D.L. (1981). "Modeling Data as a Fractional Gaussian Noise," Ph.D. Thesis, Princeton University, Department of Statistics, Princeton, New Jersey, USA.

Park, S.K. and Miller, K.W. (1988). "Random number generators: Good ones are hard to find," *Communication of the ACM*, **31**, 1192–1201.

Peitgen, H-O. and Saupe, D. eds. (1988). *The Science of Fractal Images*, Springer-Verlag, New York, NY.

Press, W.H., Teukolsky, S.A., Vetterling, W.T. and Flannery, B.P. (1997). *Numerical Recipes in C: The Art of Scientific Computing*, (2nd Ed.), Cambridge University Press, New York.

Schrage, L. (1979). "A more portable random number generator," *Assoc. Comput. Mach. Trans. Math. Software*, **5**, 132–138.

Voss, R. (1985). "Random fractal forgeries," *SIGGRAPH Conference Tutorial Notes*, ACM, New York, .

Whittle, P. (1956). "On the Variation of Yield Variance with Plot Size," *Biometrika*, **43**, 337-343.

Application of the Random Finite Element Method

Michael A. Hicks

Delft University of Technology, The Netherlands

Soil heterogeneity influences material behaviour and geotechnical performance. This chapter illustrates, through a series of numerical examples of varying degree of complexity, how the random finite element method (RFEM) may be used to assess the influence of soil heterogeneity on geotechnical performance and uncertainty. This method links random field theory for modelling the spatial variability of soil properties with finite elements for modelling geotechnical performance, within a Monte Carlo framework. It uses the soil property point and spatial statistics as input, enabling the structural "output" response to be described in terms of either reliability or probability of failure. Numerical analyses are presented which demonstrate the importance of three-dimensionality when considering the effects of heterogeneity; indeed, results demonstrate just how difficult it is to compute 2D failures in a heterogeneous material. It is also shown how RFEM provides a self-consistent framework for explaining the concept of characteristic values in Eurocode 7, as well as providing a means by which reliability-based characteristic values may be determined. Characteristic values are shown to be problem-dependent and a function of two competing factors: the spatial averaging of properties along potential failure surfaces, which reduces the coefficient of variation of the property values; and the tendency of failure mechanisms to follow the path of least resistance, which causes an apparent reduction in the property mean. The use of RFEM in assessing the liquefaction potential of sand in two underwater slope case histories is also reported.

1 Introduction

Heterogeneity is the spatial variability of soil properties and it occurs at multiple scales: at the very small scale, as seen in the arrangement of solid particles in granular soils and in the fibrous nature of organic soils such as peat; at the centimeter to meter scale, as seen in the spatial variability of soil properties within soil layers; at the medium scale, as seen in the geological layering of soils of different types; and at the very large (e.g. regional) scale. Heterogeneity influences the hydro-

mechanical behaviour of soils and the performance of geotechnical structures. Moreover, the presence of heterogeneity leads to uncertainty in ground conditions and thereby to uncertainty in design [Arn01, Hic02].

This chapter focuses on the heterogeneity that exists in so-called "uniform" soil layers, and on the influence that this heterogeneity has on geotechnical performance and uncertainty. It includes aspects of the measurement and statistical quantification of heterogeneity. In particular, it describes how heterogeneity may be modelled using random field theory, and how random fields may be linked with finite elements within a probabilistic framework to enable computations of reliability of geotechnical structures, a methodology often referred to as the random finite element method (RFEM) [Fen01]. The influence of heterogeneity on geotechnical performance is illustrated through a series of slope stability applications.

2 Basics of stochastic analysis

Figure 1(left) shows the variation in soil property X through a so-called "uniform" soil layer. In a conventional (deterministic) analysis, a single "representative" value of the soil property is adopted, and, when this is used in the analysis of a geotechnical structure, it leads to a single factor of safety. Unfortunately, however, this factor of safety tells the engineer nothing about the probability of failure.

In contrast, stochastic analysis makes use of all data from the layer, and expresses them in the form of a probability density function, as illustrated by the normal distribution in Figure 1 (right). This distribution is characterized by the mean and standard deviation of X, denoted by μ and σ, respectively, and by the coefficient of variation, expressed as $V = \sigma/\mu$. A third statistical parameter, the scale of fluctuation θ, is illustrated in Figure 1 (left). This is the distance over which soil properties are significantly correlated [Van02] and may be regarded as a function of the distance between, for example, adjacent strong or weak zones. Figure 2 illustrates the influence of θ, relative to the domain dimension D, on the spatial variation of the soil property X in two dimensions, in which dark and light zones indicate high and low values of X respectively. For small θ/D the property values change rapidly over small distances, whereas, for large θ/D, the spatial variation is much more gradual. In this figure, the spatial variability is modelled by random fields generated by local average subdivision (LAS) [Fen02]. Note that the same scale of fluctuation has been used in all directions, so these are examples of isotropic random fields, whereas, in practice, the scale of fluctuation will be larger in the horizontal plane due to the process of soil deposition.

Stochastic analysis leads to the performance of a structure being expressed in probabilistic terms, rather than in terms of a single factor of safety. For example, structure response is often described in terms of probability of failure, or in terms of reliability (which is the probability of failure not occurring). Although there are a range of approaches to conducting stochastic analysis [Hic02], this chapter focuses on the use

of the random finite element method (RFEM). Sections 2.1 to 2.3 briefly summarize three stages that may be followed in preparing for, and conducting, an RFEM analysis.

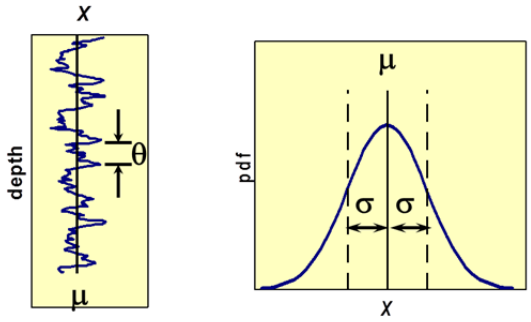


Figure 1: Illustrating the statistics of soil property X; X as a function of depth (left), probability density function of X (right) [Sam01].

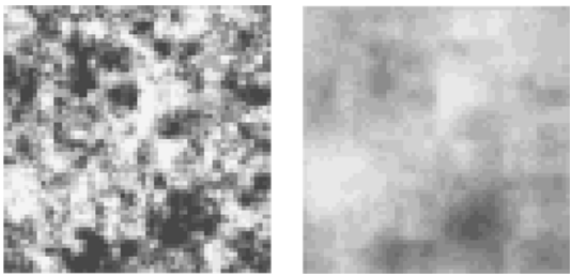


Figure 2: Two-dimensional random fields of X; $\theta/D = 0.1$ (left), $\theta/D = 1.0$ (right) [Sam01].

2.1 Pre-analysis stage: statistical site characterization

The site to be analyzed is characterized by, firstly, identifying the geological layering (i.e. sand, clay, peat, and so on), and secondly, characterizing the heterogeneity of soil properties within layers. Hence, for a given soil layer, the variation of the soil property X is represented by its point and spatial statistics.

The point statistics (μ and σ) may be determined from either in situ or laboratory data. However, in situ data are preferable, since these reduce the possibility of exaggerated estimates for the standard deviation, due, for example, to sampling or testing procedures. Furthermore, in situ data are needed anyway for determining the scale of fluctuation. In this respect, the continuous resistance profiles obtained with CPTs are particularly useful.

The following sequential process may be adopted for the soil property X [Hic08]:

- Any depth trends in the data are identified; i.e. are μ and/or σ functions of the depth z?
- Determine $\mu(z)$ and $\sigma(z)$, which may be used to define the probability density function.
- Remove the depth trend from the raw data and determine the vertical scale of fluctuation θ_v; e.g. by using the method proposed by [Wic01].
- By comparing closely-spaced de-trended property profiles, determine the horizontal scale of fluctuation θ_h. This presents the biggest problem, as sufficient data are needed to accurately determine θ [DeG01]. While this is not an issue for the vertical direction when using CPT, it has obvious implications for the required intensity of in situ testing when characterizing spatial correlations in the horizontal plane, as highlighted by [Llo01, 02].

Note that, although the above procedure will result in the characterization of a site in terms of the point and spatial statistics of a soil property, the actual heterogeneity will only be known at the locations of, for example, the CPTs. However, the statistics are needed to generate numerical predictions (i.e. random fields) of the heterogeneity across the whole site. Clearly, if it were possible to test every part of a site the heterogeneity would be known everywhere, and then there would be no need for numerical predictions of the heterogeneity and no need for a Monte Carlo simulation. Hence the need for RFEM, in assessing the performance of a structure, is not due to the soil being heterogeneous; rather, it is needed because there is incomplete knowledge about a site and therefore uncertainty about how a structure will perform.

2.2 Analysis stage: Monte Carlo simulation

For a given set of soil property statistics, for example as determined from CPT data for the site under consideration, a series of random property fields is generated and, for each, the geotechnical problem is analyzed to give a range of solutions. Each random field generation and subsequent finite element analysis of the structure, using that random field, is known as a realization. Hence, RFEM involves multiple realizations as part of a Monte Carlo simulation. For all realizations, the random fields will look similar, as they will all have been generated using the same set of statistics, but they will all be different with respect to the distribution of strong and weak zones, and each random field will lead to a different solution when used in a finite element analysis. Figure 3 shows four random fields generated using the same set of input (soil property) statistics, while Figure 4 shows computed shear strain invariant contours at failure for four slopes based on the same input statistics of soil shear strength.

Each realization in the Monte Carlo simulation is a standard deterministic finite element analysis [Smi01], but with each element in the finite element mesh being assigned a different value of X that is mapped onto the mesh from the random field. Indeed, it is also possible to assign a different value of X to every element sampling

(integration) point, in order to optimize the level of heterogeneity that may be modelled for a given level of finite element discretization. However, whether the random field is mapped onto the finite element mesh at the element or sampling point level, an approximation is involved, since X is thereby assumed to be constant over each element domain (or over that part of the element associated with each sampling point): that is, the field is discrete rather than continuous. The aim, therefore, is to generate a discrete random field in which the point statistics are adjusted (to account for the finite size of an element, or sampling point "domain") so that they are equivalent to those of the underlying continuous field. For all applications presented herein, the random fields have been generated using local average subdivision [Fen02], so-named because of its use of local averaging theory.

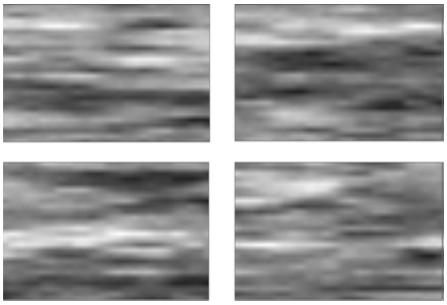


Figure 3: Four random fields of X based on the same input statistics [Hic08].

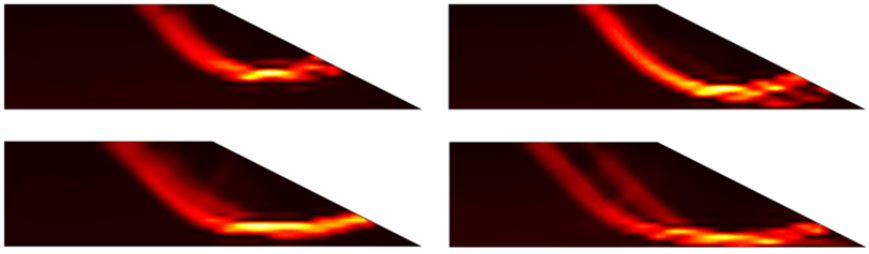


Figure 4: Four slope failure mechanisms based on the same input statistics [Hic08].

In simple terms, for each realization: (a) a discrete random field of X is generated, based on the statistics of X (μ, σ, θ_v, θ_h); (b) the random field is mapped onto the finite element mesh; (c) the problem is analyzed by finite elements. However, many problems involve the heterogeneity of more than one material property: that is, $\mathbf{X} = \{X_1\ X_2\ X_3\ \ldots\ X_n\}^T$, in which n is the number of variables. In such cases, two approaches are possible: (a) the multivariate approach is to generate a separate random field for each variable, and to cross-correlate between fields to account for parameter inter-dependency ([Fen03], [Arn02]), for example, higher friction angles are likely to be associated with an increased tendency for dilation; (b) the reduced-

variate approach is to generate random fields for smaller numbers of parameters (e.g. relative density) from which other parameters (e.g. friction angle) may be back-figured; for example, [Pop01] generated cross-correlated, bi-variate, random fields of cone tip resistance and soil classification index, while [Hic06] generated univariate random fields of state parameter [Bee01].

2.3 Post-analysis stage: measures of structure performance

The results of the Monte Carlo realizations are typically presented in the form of a "performance" probability density function (PDF) or cumulative distribution function (CDF). The probability of failure (or of failure not occurring) can then be found, either by proportioning the area under the PDF or by reading directly from the CDF. The Monto Carlo simulation continues until the output statistics of, for example, factor of safety, reach an acceptable level of convergence. Note that the required number of realizations in an RFEM analysis will generally be substantially less than for a simple probabilistic analysis based only on the point statistics (i.e. 100s rather than 1000s), due to the spatial averaging of property values. Indeed, in the limit when θ approaches zero, only one realization would be required.

3 Influence of heterogeneity on slope reliability

This section investigates the influence of heterogeneity of undrained shear strength on the reliability of slopes that are long in the third dimension. Firstly, a simple 2D investigation is described, to illustrate some of the basic characteristics of analyses involving soil heterogeneity. Next, the influence of heterogeneity on slope failure in three-dimensions is highlighted. Finally, the implications for assessing the stability of dykes and embankments is illustrated and discussed. In each case, the undrained shear strength is represented by a truncated normal distribution to avoid the possibility of negative values. This is a reasonable distribution for this soil property, due to coefficient of variation generally lying in the range $0.0 < V < 0.3$ (that is, the very small proportion of values that will need to be truncated will have a negligible influence on the analysis) [Hic07].

3.1 Stochastic analysis of 2D slope reliability

Figure 5 shows the finite element mesh used to model a 1:2 slope characterized by a spatially varying undrained shear strength c_u [Hic08]. The height of the slope is 10 m, the volumetric weight of the soil is $\gamma = 20$ kN/m^3, and the statistics of c_u are a mean that increases linearly with depth, from 10 kPa at the top boundary to 50 kPa at the bottom boundary, and a constant coefficient of variation of 0.3. The vertical scale of fluctuation is $\theta_v = 1.0$ m, whereas various horizontal scales of fluctuation have been considered [Hic08]. Figure 6 shows a typical ranfom field of c_u for $\xi =$

$\theta_h/\theta_v = 12$, in which the darker zones indicate higher values of c_u. The soil has been modelled by a Tresca failure surface and the following elastic properties: Young's modulus, $E = 100{,}000$ kPa, and Poisson's ratio, $\nu = 0.3$.

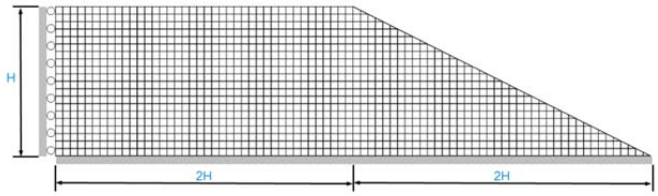


Figure 5: Geometry, boundary conditions and finite element mesh [Hic08].

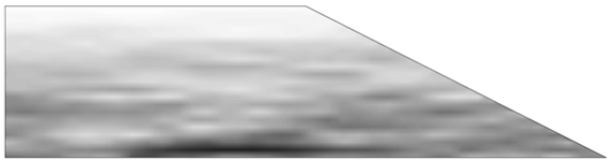


Figure 6: Typical random field for $\xi = 12$ [Hic08].

The slope has been analyzed using the strength reduction method. For each random field, gravitational loading is applied to generate the in situ stresses in the slope and the crest settlement, Δ, due to the soil self-weight, is recorded. The slope is then repeatedly analyzed for progressively weaker soil profiles until the slope fails, as indicated by a sudden increase in the crest settlement. For each re-analysis, the random field of c_u is the original random field for the realization generated by the input statistics, scaled down by a factor F. The scaling factor that causes failure is the factor of safety of the slope (based on the original random field).

Figure 7 shows that, when no heterogeneity is considered, the factor of safety is close to the analytical solution of $F = 1.6$. However, Figure 8 shows that, when heterogeneity is considered, there is a wide range of possible solutions. Moreover, the mean factor of safety is significantly less than the deterministic solution based on the underlying depth-dependent mean. This is because failure mechanisms follow the path of least resistance: that is, they are attracted to the weaker zones and try to avoid (where possible) the stronger zones.

[Hic08] considered several values of ξ for this boundary value problem and showed that the distribution of factor of safety tended to converge for values of ξ that were lower than likely to be encountered on site. This implied that θ_h need not always be accurately known, since a conservative solution could be found by assuming $\xi = \infty$. However, later analyses in 3D, summarized in Section 3.2, have suggested that a more accurate knowledge of θ_h may be important. [Hic07] carried out a detailed 2D

investigation to illustrate the importance of accounting for both the anisotropy of the heterogeneity ($\xi > 1$) and the depth-dependency of the underlying mean c_u, whereas [Hic09] investigated the influence of slope angle.

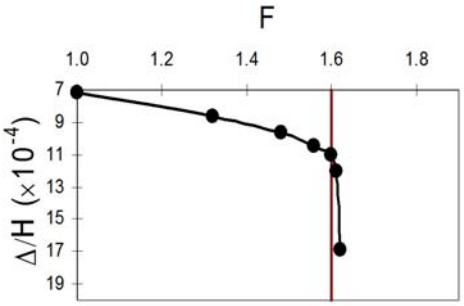


Figure 7: Mobilized safety factor versus crest settlement (deterministic solution) [Hic08].

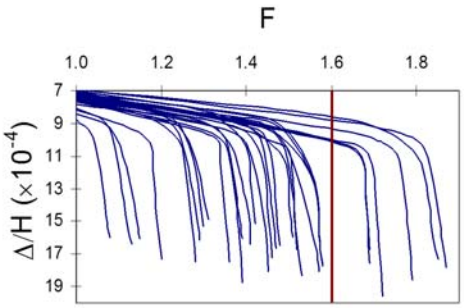


Figure 8: Mobilized safety factor versus crest settlement (stochastic solution, $\xi = \infty$) [Hic08].

3.2 Influence of heterogeneity on 3D slope reliability

[Spe01, 02] and [Hic10] carried out 3D RFEM analyses for a long slope in a soil characterized by a spatially varying undrained shear strength with constant (i.e. depth-independent) mean and coefficient of variation. Figure 9 shows the problem geometry and mesh details. The 1:1 slope is H = 5 m high and L = 100 m long, and is modelled using 8000 20-node brick elements with 2 × 2 × 2 Gaussian integration.

As in Section 3.1, the slope has been loaded by applying gravity loading to generate the in situ stresses. However, rather than analyzing the slope for a given set of statistics and finding the distribution of factor of safety, this investigation focusses on finding the relationship between reliability and factor of safety based on the mean c_u.

The process starts with $\mu_{F=1.0}$, which is the mean c_u at which the slope would just start to fail if there were no spatial variation in c_u. Therefore, for a given value of F and ξ, the point and spatial statistics are: $\mu = \mu_{F=1.0} \times F$; $\sigma = \mu \times V$; $\theta_h = \theta_v \times \xi$, where, for this investigation, $\mu_{F=1.0} = 16.1$ kPa, $V = 0.3$ and $\theta_v = 1.0$ m. These statistics are used to generate N random fields of c_u, and, for each realization, the slope is analyzed by finite elements. The percentage reliability is then given by $R = (1 - (N_f/N)) \times 100$, where N_f is the number of realizations in which the slope fails under its own self weight.

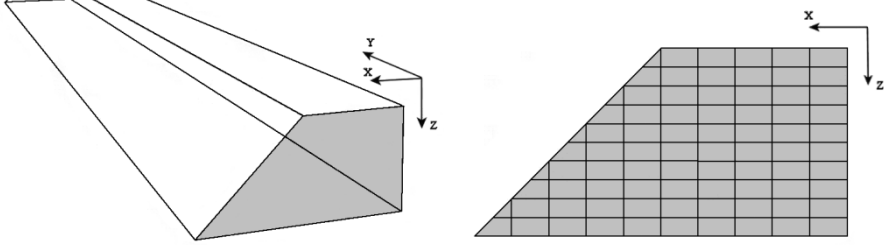


Figure 9: Isometric projection of 3D slope and cross-section through mesh [Spe02, Hic10].

Figure 10 shows the relationship between reliability and global factor of safety for a 2D (i.e. plane strain) analysis. At a factor of safety of 1.0 (based on the mean c_u) $R < 50\%$ for all values of ξ, due to failure being attracted to the weaker zones. It is clear that, although the solution is dependent on the horizontal scale of fluctuation, the solution has converged for $\xi > 6$ in this example.

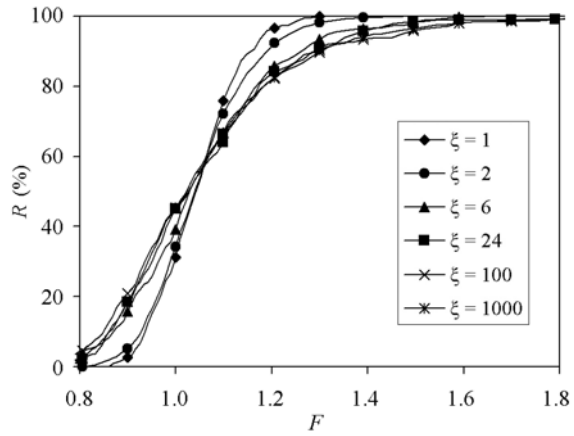


Figure 10: Influence of degree of anisotropy of the heterogeneity on reliability versus global factor of safety (2D analysis) [Spe02, Hic10].

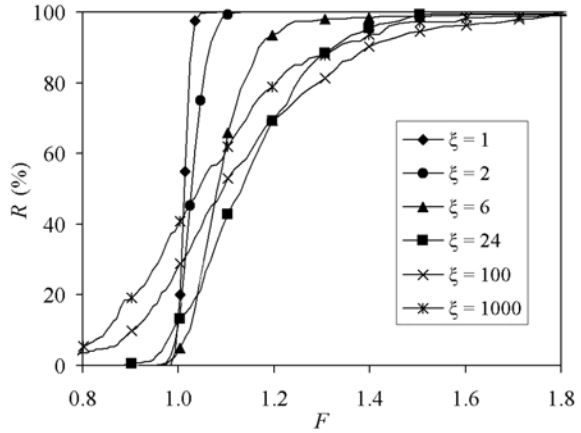


Figure 11: Influence of degree of anisotropy of the heterogeneity on reliability versus global factor of safety (3D analysis) [Spe02, Hic10].

Figure 11 shows the results of the equivalent 3D analyses. In evaluating these results, [Spe02, Hic10] identified three categories of failure mode, which are illustrated by the typical deformed meshes in Figure 12. The failure modes are as follows:

- Mode 1: For $\theta_h < H$, the scale of fluctuation is relatively small in all directions and it becomes harder for failure to propagate through semi-continuous weaker zones. In particular, for very small values of θ_h, the failure mechanism passes through strong and weak zones in almost equal measure. There is, therefore, considerable averaging of soil properties over potential failure planes, and the soil layer behaves like a homogeneous soil characterised by the mean c_u. This theory is supported by R increasing from 0-100% as F passes through 1.0 (Figure 11), and by failure originating from the slope toe and extending along the entire length of the slope (Figure 12 (left)).
- Mode 2: For $H < \theta_h < L/2$, it is possible for failure to propagate through semi-continuous weaker zones, which results in discrete (3D) failure mechanisms (Figure 12 (centre)). In this case, R is a function of slope length, since, as the slope becomes longer, there is an increased possibility of encountering a zone weak enough to trigger failure.
- Mode 3: For $\theta_h > L/2$, there is an increased likelihood of the failure mechanism extending along the length of the slope (Figure 12 (right)), as in Mode 1. However, in contrast to Mode 1, there is now a large range of possible solutions, due to the depth of the failure mechanism being influenced by the distribution of strong and weak "sub-layers". In this case, the R versus F relationship approaches that obtained for the 2D stochastic analysis.

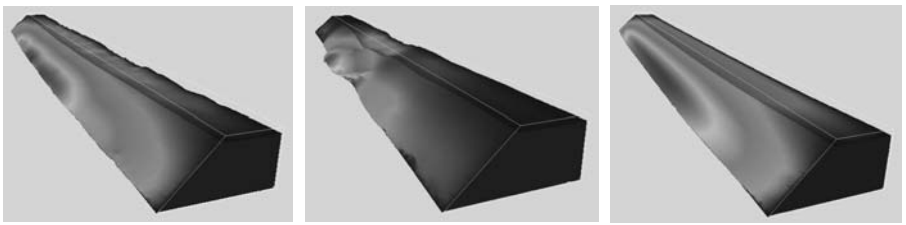


Figure 12: Typical deformed meshes and contours of out-of-face displacement for various 3D mechanisms (left to right: Mode 1, Mode 2, Mode 3) [Spe02, Hic10].

A practical implication of the results in Figure 11 is that, for Modes 1 and 3, the solution is independent of the slope length, since the failure mechanism is two-dimensional. In contrast, the solution for Mode 2 is length dependent, because the failure mechanism is three-dimensional. This has practical implications, since failures are generally three-dimensional and it is clearly impractical to analyze very long slopes (e.g. dykes and embankments) in 3D. However, [Hic10] carried out a detailed stochastic analysis of a 50 m long embankment, and then combined the results with simple probability theory to successfully predict the behavior of longer embankments that had also been analyzed using 3D RFEM.

3.3 Influence of heterogeneity on failure consequence

The investigation in Section 3.2 was extended by [Nut01, Hic11], who implemented a simple numerical scheme for automatically computing slide geometries in 3D RFEM simulations. Figure 13 shows the results for a similar slope to that analyzed in [Hic10], except that, in this case, $V = 0.2$ and a Von Mises failure criterion has been adopted. The figure shows the influence of the horizontal scale of fluctuation on reliability versus global factor of safety, as well as on failure volumes and lengths for individual realizations (which have been expressed as percentages of the total mesh volume and length, respectively). Because the same slope geometry and θ_v as used in [Hic10] have been adopted, the values of ξ in Figure 13 are directly comparable with those in Figure 11.

Figure 13(a) shows that, when $\theta_h = H/5$, the slide volumes and lengths are consistent with those obtained when the slope fails along its entire length (indicating Mode 1 failure). In contrast, Figure 13(b) shows that, when $\theta_h = 1.2H$, there is a wide range of slide geometries (indicating Mode 2 failure). Similarly, Figures 13(c) and 13(d), corresponding to $\theta_h \approx L/8$ and $\theta_h \approx L/2$, respectively, indicate Mode 2 failure. Although Figure 13(e) does reveal an increase in the number of larger slides for $\theta_h = L$, suggesting some Mode 3 failures, it is apparent that most slides are still Mode 2. Indeed, although Figure 13(f) shows mainly Mode 3 failures for $\theta_h = \infty$, [Nut01, Hic11] demonstrated just how difficult it is to compute 2D slope failures in a soil that is heterogeneous, and presented results for a slope with a foundation layer which showed an even greater tendency for Mode 2 failures.

The ability to automatically compute slide volumes is an important first step towards benchmarking simpler analytical and probabilistic models used in design [Li01, 02]. This is because there is a need to quantify slide geometries when comparing with methods based on predefined (e.g. cylindrical) failure mechanisms [Cal01, Van01].

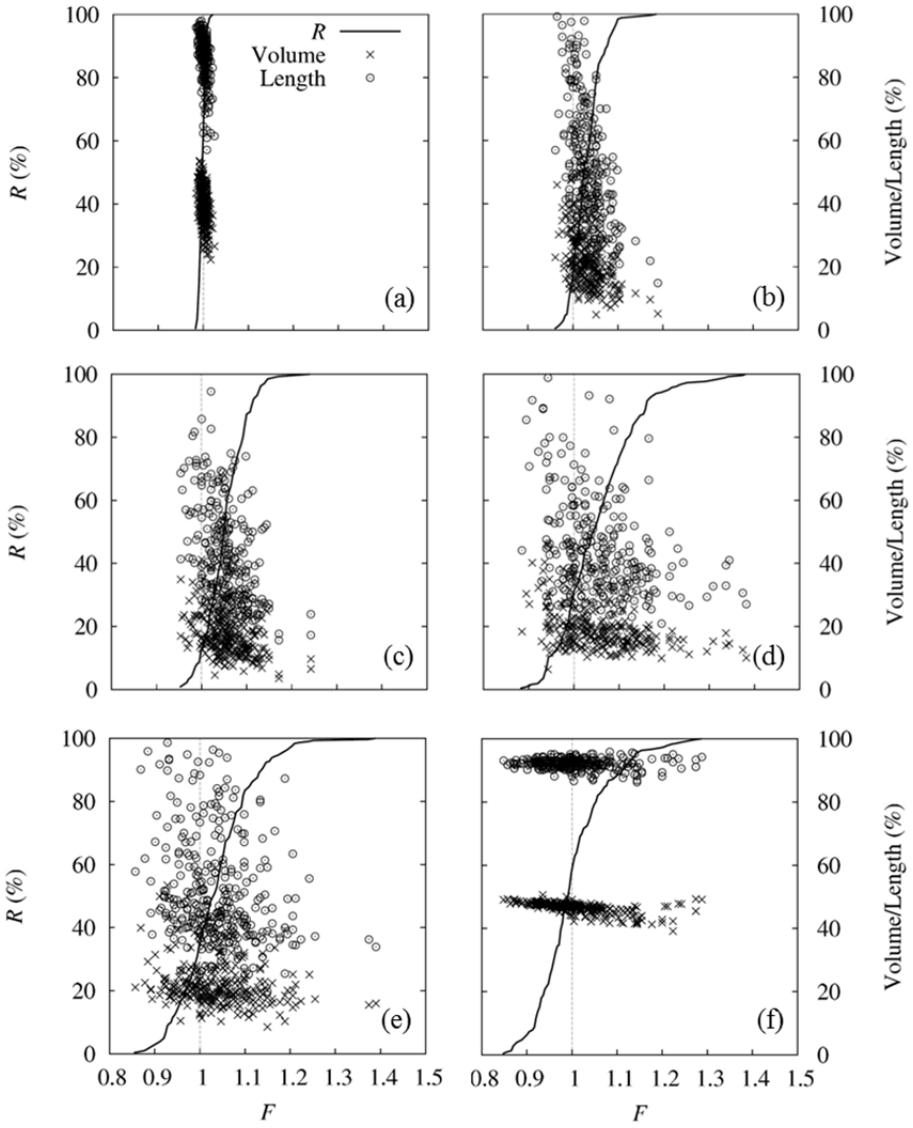


Figure 13: Influence of ξ on slide volume and length for a 3D slope; (a) $\xi = 1$, (b) $\xi = 6$, (c) $\xi = 12$, (d) $\xi = 48$, (e) $\xi = 100$, (f) $\xi = \infty$ [Nut01, Hic11].

4 Stochastic explanation of characteristic values

This section considers the issue of heterogeneity within the context of characteristic soil property values advocated in Eurocode 7 (EC7) [CEN01]. It is shown that stochastic analysis may be used as an aid to understanding the philosophy and nature of characteristic values, as well as providing a framework for deriving reliability-based characteristic values in line with EC7 [Hic03, 05, 08].

4.1 Extracts from Eurocode 7

The importance of accounting for the variability of soils is highlighted in Section 2.4.5.2 of EC7, "Characteristic values of geotechnical parameters" [CEN01]. Table 1 lists some of the main clauses, including: Clause (4)P, which highlights the spatial nature of soil variability, the uncertainty this causes and the problem-dependency of characteristic values; Clause (7), which emphasizes the importance of the mean over the domain of influence; Clause (8), which considers the special case of local failure; and Clause (11), which considers the use of statistical methods.

[Hic03] gave a detailed review of Section 2.4.5.2 by explaining selected clauses, clarifying the relationship between clauses and addressing areas of potential confusion. In particular, the paper focused on the statistical definition of a characteristic value given in Clause (11) and explained how it is, despite first appearances, completely consistent with Section 2.4.5.2 as a whole, including Clauses (7) and (8) and the footnote to Clause (11).

Clause (11) states that "the characteristic value should be derived such that the calculated probability of a worse value governing the occurrence of the limit state under consideration is not greater than 5%". This implies a minimum level of reliability of 95% regarding the response of the structure (before application of partial safety factors), and appears to contradict Clauses (7) and (8) and the footnote to Clause (11) which focus on property values rather than structure response. However, [Hic03] used Figure 14 to demonstrate that the latter are merely special cases of Clause (11).

4.2 Reliability-based characteristic values

Figure 14 (top) shows the probability density function of a material property X, which, to simplify the illustration, is assumed to be normal with a mean value X_m. The simplest way to derive a reliability-based characteristic value X_k is to proportion the area under the distribution as indicated. However, this is not consistent with Clause (11), as it merely defines a value of X_k for which there is a 95% probability of a larger value.

Table 1: Extracts from Section 2.4.5.2 of Eurocode 7 [CEN01, Hic05].

No.	Clause
(4)P	The selection of characteristic values for geotechnical parameters shall take account of the following: ● geological and other background information, such as data from previous projects; ● the variability of measured property values and other relevant information, e.g. from existing knowledge; ● the extent of the field and laboratory investigation; ● the type and number of samples; ● the extent of the zone of ground governing the behaviour of the geotechnical structure at the limit state being considered; ● the ability of the geotechnical structure to transfer loads from weak to strong zones in the ground.
(7)	The zone of ground governing the behaviour of a geotechnical structure at a limit state is usually much larger than a test sample or the zone of ground affected in an in situ test. Consequently the value of the governing parameter is often the mean of the range of values covering a large surface or volume of the ground. The characteristic value should be a cautious estimate of this mean value.
(8)	If the behaviour of the geotechnical structure at the limit state considered is governed by the lowest or highest value of the ground property, the characteristic value should be a cautious estimate of the lowest or highest value occurring in the zone governing the behaviour.
(11)	If statistical methods are used, the characteristic value should be derived such that the calculated probability of a worse value governing the occurrence of the limit state under consideration is not greater than 5%. NOTE: In this respect, a cautious estimate of the mean value is a selection of the mean value of the limited set of geotechnical parameter values, with a confidence level of 95%; where local failure is concerned, a cautious estimate of the low value is a 5% fractile.

Figure 14 (bottom) gives a more general derivation of X_k that is consistent with Clause (11) and, by association, with all other clauses in Section 2.4.5.2. This involves proportioning the area under a modified distribution of X that has been backfigured from the geotechnical response of the structure itself. The modified distribution is narrower than the underlying property distribution due to the averaging of property values over potential failure surfaces. It is also shifted to the left, due to the tendency for failure to propagate through weaker zones. Hence, although it may be reasonable to take a conservative estimate of the mean property value over a potential failure surface as the characteristic value for that mechanism, this mean will generally be smaller than the mean of the underlying property distribution. Variance

reduction methods may therefore give an unsafe solution if no account is taken of the reduction in the mean.

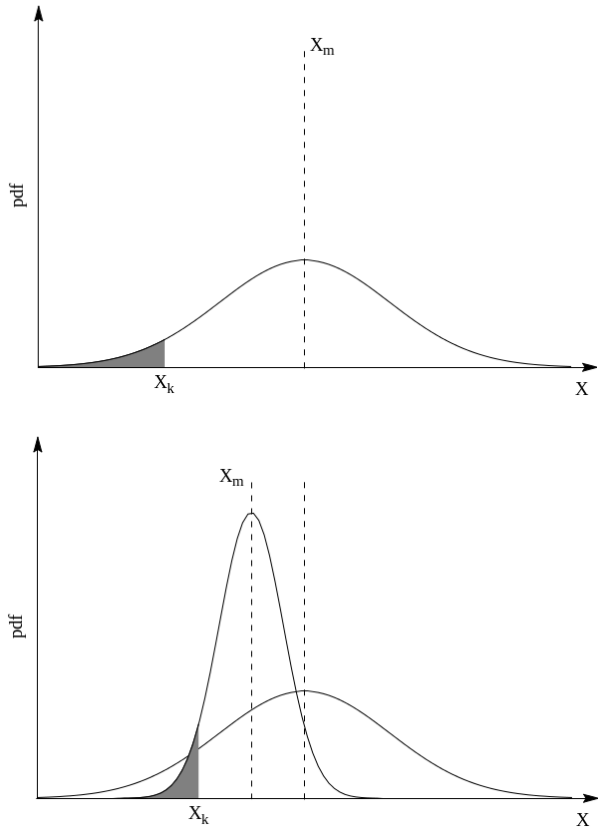


Figure 14: Derivation of characteristic property values satisfying Eurocode 7; basic definition of X_k (top), general definition of X_k (bottom) [Hic03].

[Hic03] explained how the modified property distribution is a function of the underlying distribution, the spatial correlation of property values, the problem being analyzed and the quality and extent of site investigation data. Moreover, the modified distribution has two limits:

- When the spatial scale of fluctuation is very small relative to the problem domain there is much averaging of soil properties, so that the standard deviation approaches zero and the mean tends to the mean of the underlying distribution. In this case a cautious estimate of the mean is appropriate, as advocated by Clause (7) and the first part of the footnote to Clause (11).
- When the spatial scale of fluctuation is very large relative to the problem domain there is a very large range of possible solutions, so that the modi-

fied distribution approaches the underlying distribution. In this case Clause (8) and the second part of the footnote to Clause (11) are relvant.

[Hic03] also explained how the modified property distribution in Figure 14 (bottom) may be derived for general problems, based on earlier work using RFEM by [Hic08], while [Hic05] extended this earlier work by illustrating the process for a 3D slope.

4.3 Derivation of characteristic values using RFEM

[Hic05] analyzed the slope shown in Figure 9, using the same finite element mesh, to illustrate the derivation of reliability-based characteristic values. In this case, an elastic, perfectly plastic Von Mises model was used, and the statistics of undrained shear strength were: $\mu = 40$ kPa, $\sigma = 8$ kPa, $\theta_v = 1.0$ m; $1.0 < \theta_h < 1000.0$ m.

For each value of θ_h a Monte Carlo simulation was carried out, comprising 250 realizations, with each realization comprising the following steps: (a) the generation of a random field of c_u; (b) the finite element analysis of the slope by applying gravity loading to generate the in situ stresses, assuming a soil unit weight of 20 kN/m^3. The second step involved the strength reduction method for computing the factor of safety F of the slope, in the same manner as described in Section 3.1 [Hic08]. Hence, each Monte Carlo simulation resulted in a distribution of factor of safety, from which a distribution of "equivalent" values of undrained shear strength were back-figured. This was achieved through relating F and c_u via the slope stability number [Tay01].

Figure 15 shows the distributions of F (top) and equivalent c_u (bottom) for $\theta_h/\theta_v = 48$. The latter distribution is directly comparable to the equivalent distribution in Figure 14 (bottom), and so may be used to determine reliability-based characteristic values for c_u. Figure 16 summarizes the results of all analyses, by plotting the mean and standard deviation of the equivalent c_u as a function of θ_h/θ_v and comparing these results with the underlying (i.e. input) mean and standard deviation. For $\theta_h/\theta_v = 1$, the prevalence of Mode1 failures is reflected by a mean equivalent c_u approaching 40 kPa and a standard deviation approaching zero. For larger θ_h/θ_v, Mode 2 failures prevail, resulting in a decrease in the mean (due to the greater relative influence of the weaker zones) and an increase in the standard deviation. Finally, as θ_h exceeds about half the length of the slope, as represented by $\theta_h/\theta_v = 50$, the tendency for discrete failures reduces (albeit gradually, as indicated in Figure 13) due to the influence of the mesh boundaries. Hence, at the same time there is an increasing tendency for Mode 3 failures initiating at depths influenced by the distribution of weak and strong "layers". This results in the mean approaching the underlying mean of 40 kPa. However, even though the standard deviation reaches a maximum of around 3 kPa, it remains well below the underlying standard deviation of 8 kPa due to the small value of θ_v causing considerable averaging of property values in the vertical direction.

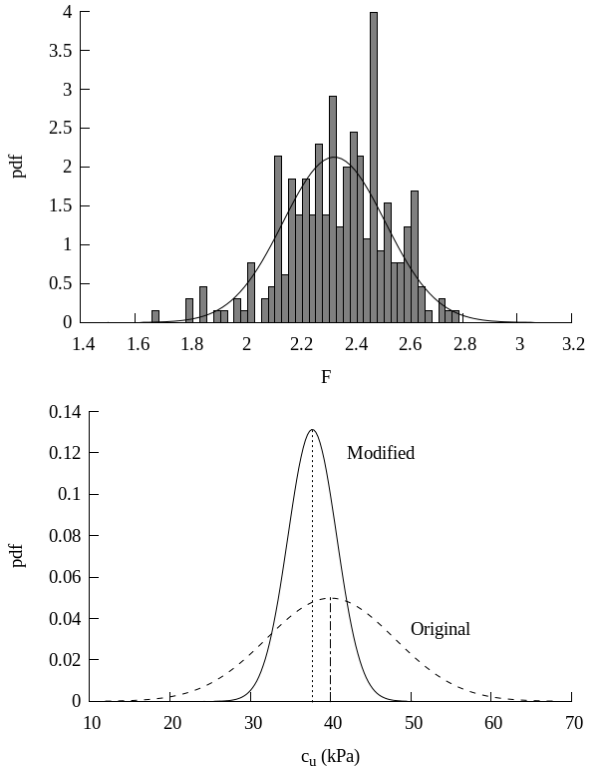


Figure 15: Distributions of factor of safety (top), and equivalent undrained shear strength (bottom) [Hic03].

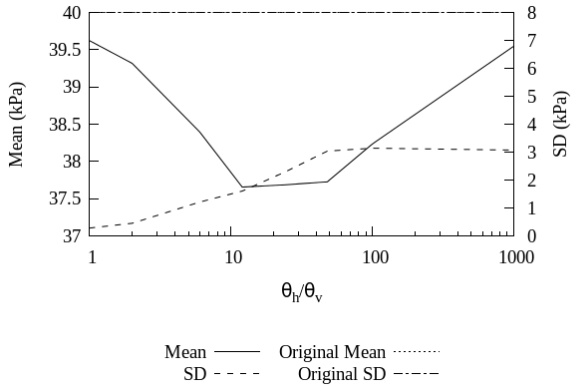


Figure 16: Influence of θ_h/θ_v on the mean and standard deviation of the "equivalent" c_u [Hic03].

5 Influence of heterogeneity on liquefaction potential

This section uses RFEM to investigate the influence of sand heterogeneity in two slope liquefaction case histories. In these analyses, the challenge is to model the spatial variability of sand relative density and the impact this has on saturated slope stability under rapid loading. Hence this requires a sophisticated soil model and, by implication, the spatial variation of numerous cross-correlated material parameters over the problem domain. For modelling the liquefaction potential of a site subjected to seismic loading, [Fen03] adopted a multi-variate random field approach, whereas [Pop01] generated bi-variate random fields of relative density and soil classification index, from which soil property values were back-figured, in order to assess the liquefaction potential of an artificial island sand core due to cyclic ice loading.

[Bak01, Hic06, Oni01, Won01] adopted a similar "reduced-variate" approach to [Pop01], except that they based their investigations on generating univariate random fields of state parameter [Bee01]. The state parameter at a given mean effective stress is given by $\psi = e\text{-}e_{cs}$, where e and e_{cs} are the current and critical state void ratios, respectively. Hence, positive and negative values of ψ indicate loose and dense soils. As a rough guide, the following categories of sand state may be suggested: (a) dense to medium dense, $-0.20 < \psi < -0.10$; (b) mildly dense, $-0.10 < \psi < -0.05$; (c) loose, $-0.05 < \psi < 0.00$; (d) very loose, $\psi > 0.00$. Figure 17 illustrates the undrained triaxial compression behavior of a sand in different states, in which t and s represent the deviatoric and mean effective stress invariants.

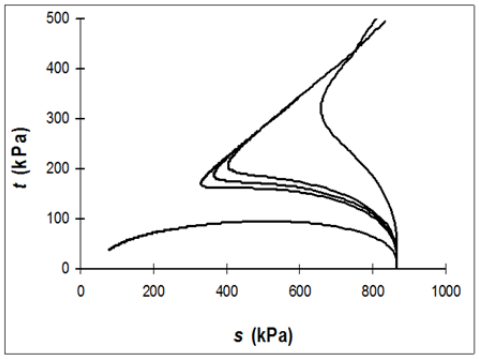


Figure 17: Influence of sand state on undrained triaxial compression effective stress paths (ψ = +0.06, -0.05, -0.08, -0.12, -0.17) [Hic06].

In the investigations of [Bak01, Hic06, Oni01, Won01], the Monot double-hardening soil model was used to model the sand behavior [Hic01, Mol01]. Five material parameters were calibrated, of which four were state parameter dependent: these related to the peak friction angle, and the stiffness of the elastic and two plastic model components. The fifth parameter was the friction angle corresponding to no-volume change during shearing, which was assumed to be independent of sand state.

5.1 Case history 1: Nerlerk berm

The Nerlerk berm was designed to be part of a bottom-founded offshore oil exploration platform in the frozen Canadian Beaufort Sea [Hic04]. Although it was one of numerous artificial islands constructed during the 1980s in this region, it was unusual, in that it was constructed in a greater water depth than had previously been attempted at other sites: specifically, the water depth was 46 m and the design height of the berm was 36 m. This led to two non-standard design decisions. Firstly, in order to reduce construction time and transport costs, a local borrow source was used to construct the berm. Secondly, there was no equipment available that could remove the weak surficial clay layer from the seabed at a depth of 46 m. Hence the island construction went ahead using a fill of lesser quality than had been used at previous island locations and, also in contrast to other sites, the weak surficial clay was not removed prior to the berm construction in 1982. Shortly after the start of the second construction season in 1983, when the berm had reached a height of 26 m, the structure experienced a series of liquefaction slides and the project was eventually abandoned with the loss of more than $ 100 million.

[Hic04] investigated possible trigger mechanisms for the liquefaction slides, and concluded that the most likely cause of failure was a combination of limited movement in the clay layer triggering liquefaction near the berm crest during rapid loading (due to the berm construction). This investigation had assumed the fill to be in a loose and liquefiable state, even though CPT data from the site had indicated a mainly denser material. Hence [Hic06] used RFEM to investigate whether it was possible for pockets of very loose material to dominate the stability of a predominantly dilative sand fill. They calibrated the Monot soil model against 74 drained and undrained triaxial compression tests on Erksak sand, for a wide range of sand states, while state parameter statistics were derived from 71 CPTs from two Beaufort Sea sand islands. Based on these statistics and the soil model calibration, multiple realizations of the berm under rapid loading were analyzed.

Figures 18 and 19 (top) show typical realizations of state parameter for a cross-section taken through the upper half of the Nerlerk berm, for both isotropic and anisotropic spatial variability, in which the darker zones indicate the sand in a denser state. In both cases, θ_v = 1.0 m, whereas, for the anisotropic case, the scale of fluctuation along the line of the slope is assumed to be eight times larger than θ_v. This value was chosen based on evidence from closely spaced CPTs at another island location [Oni01, Won01].

Figures 18 and 19 (bottom) show typical contours of shear strain invariant that developed in the slope for the two types of random field, when the slope was loaded by increasing gravity with the soil in an undrained state. For the isotropic case (Figure 18) although a rather complicated “spider’s web” typed of mechanism is observed, the strains are very small, due to the stronger zones compensating for the weaker zones and holding the structure together in a stable state. [Hic06] conducted multiple realizations for isotropic spatial variability and showed that, in all cases, the re-

sponse of the structure was very close to the deterministic response computed for the slope based on the mean state parameter.

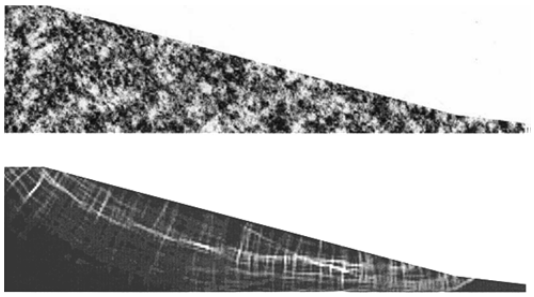


Figure 18: Typical random field of state parameter (top) and deformation mechanism (bottom) for isotropic heterogeneity [Hic06]

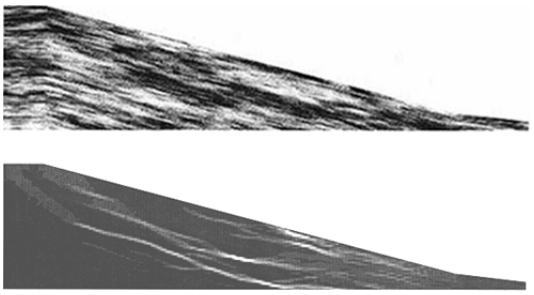


Figure 19: Typical random field of state parameter (top) and deformation mechanism (bottom) for anisotropic heterogeneity [Hic06]

In contrast, for the much more realistic anisotropic case (Figure 19), the strains are significantly larger and the slope fails due to liquefaction through semi-continuous weaker zones arising from deposition-induced anisotropy. [Hic06] showed that, in this case, a deterministic analysis based on the mean state parameter gave an upper bound solution. All analyses based on a heterogeneous sand gave a weaker response, with the lower bound solution appearing to be around the same as that obtained with a deterministic analysis using a state parameter equal to the mean minus two standard deviations. Hence, in some realizations it was the weakest material that dominated the stability of the entire structure.

[Won01] extended the investigation of Nerlerk berm by considering the whole structure (including the clay layer) and accounted for the impact of zoning within the sand fill due to the method of construction. It was shown that larger-scale spatial fluctuations due to zoning were more influential on structure response than smaller-scale fluctuations normally associated with more uniform soil deposits.

5.2 Case history 2: Jamuna Bridge

[Bak01] investigated the liquefaction of riverbed deposits during the construction of a bridge over the Jamuna River in the 1990s. The Jamuna River is a shifting braided river, in that it can take the form of multiple channels that rapidly change position from one rainfall season to the next. In order to control the lateral movement of the river during the bridge's construction, it was decided to construct two guide bunds, one on either side of the main river channel. Each bund was designed to be 30 m high with side slopes of 1:3.5-1:5, but, during 1995-96, there were around 30 liquefaction slides. The side slopes were then redesigned to 1:6 and the bridge was eventually completed in 1999.

[Bak01] adopted a similar approach to [Hic06] to investigate the influence of heterogeneity on the liquefaction potential of the riverbed deposits. A total of 22 CPTs were evaluated to determine the state parameter statistics and the Monot soil model parameters were related to state parameter through calibration against a laboratory database. Figure 20 shows a typical CPT profile, including tip resistance (extreme left) and state parameter (extreme right). The state parameter from 5-35 m depth has been de-trended with respect to the mean, and then normalized to give the probability density function shown in Figure 21. The figure shows the ability of different theoretical distributions to approximate the data. For this CPT, the best fit has been obtained with a normal distribution.

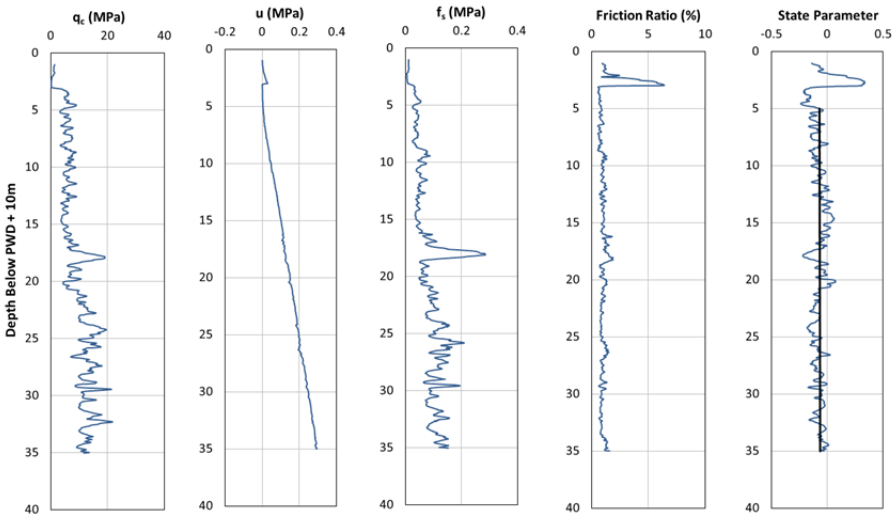


Figure 20: Data from CPT C330W [Bak01].

The mean trend line for state parameter in Figure 20 indicates that the mean is almost constant with depth. It is approximately -0.07, indicating a mildly dilating sand. However, the standard deviation is 0.05, suggesting that the weaker zones in

the deposit will be very loose and liquefiable. The vertical scale of fluctuation for this CPT profile is 1.0 m, which, from the author's experience, is on the high side for a uniform sand. However, the slightly bi-modal appearance of the PDF in Figure 21 suggests that there might be a small degree of zoning in the soil, and this is supported by Figure 20 which suggests a slightly denser zone at around 15 m depth and a slightly looser zone at depths greater than 23 m. In other words, the slightly larger θ_v is due to it being the sum of two components: a small θ_v that is associated with heterogeneity within so-called uniform soils, plus a larger θ_v due to zoning.

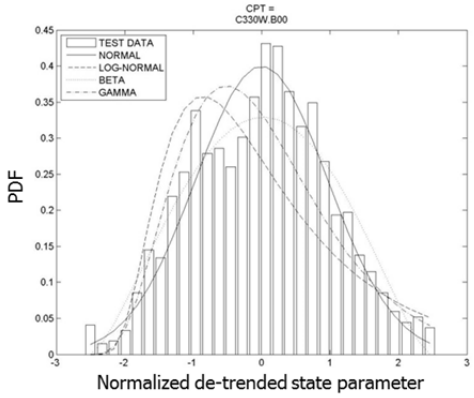


Figure 21: Probability density function of normalized de-trended state parameter for CPT C330W [Bak01].

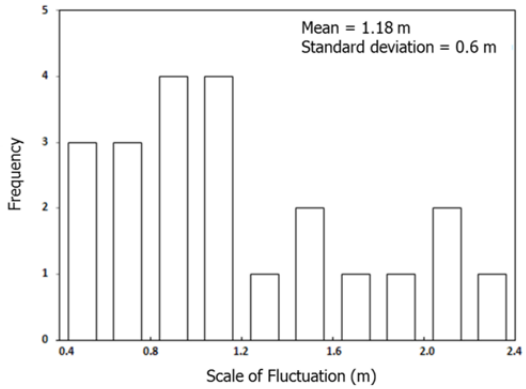


Figure 22: Frequency diagram of vertical scale of fluctuation from all CPTs [Bak01].

Figure 22 shows that the vertical scale of fluctuation at the site varies considerably from one CPT to another. In general, the following observations have been made: for $\theta_v < 0.8$ m the soil is relatively uniform, as will be apparent by a mono-modal

property distribution; for $0.8 < \theta_v < 1.2$ m there is likely to be mild to moderate zoning, as indicated by a slightly to moderate bi-modal property distribution; for $\theta_v >$ 1.2 m there is likely to be significant zoning, which will be obvious by a significantly bi-modal PDF. [Bak01] considered a plan view of the west guide bund, from which the CPTs had been taken, and showed that, while the mean state parameter did not vary that much across the site, higher values of θ_v were generally associated with those parts of the bund in which slope liquefaction occurred, thereby suggesting that the potential for liquefaction is greater for larger scales of fluctuation.

The influence of heterogeneity was investigated for a 2D cross-section through the slope of the west guide bund [Bak01]. Firstly, Figure 23 shows the results of a series of deterministic undrained analyses based on uniform soil properties, where the slope was loaded by increasing gravity. It is clear that, for very loose and very slightly dilatant soils, as represented by positive and small negative values of state parameter respectively, the slope can fail under undrained loading. In contrast, for mildly and strongly dilatant soils, failure does not occur.

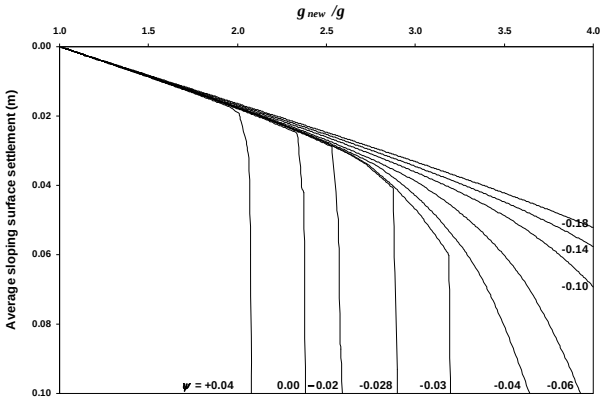


Figure 23: Undrained slope stability based on mean properties [Bak01].

Figure 24 shows the results of a small number of realizations for the same cross-section, assuming a heterogeneous sand with the same state parameter statistics as obtained for CPT C330W (Figure 21); that is, μ = -0.07, σ = 0.05, θ_v = 1.0 m. A horizontal scale of fluctuation of 8.0 m was used, in line with earlier work by [Hic06] who also used θ_h/θ_v = 8. Indeed, similar conconclusions to those obtained for the Nerlerk berm by [Hic06] were found: (a) the computed deterministic response based only on the mean state parameter gives an upper bound solution; (b) in many analyses, slope stability is dictated by the behaviour of the weakest material. Figure 24 compares the computed results with the deterministic solutions based on ψ = +0.04, 0.0 and -0.07. It is seen that the weakest response corresponds to a deterministic analysis based on $\psi \approx \mu - 1.4\sigma$. However, this is based on a small

number of realizations, so it may be that, for a larger number relaizations, $\psi \approx \mu - 2\sigma$ is more realistic as a lower bound.

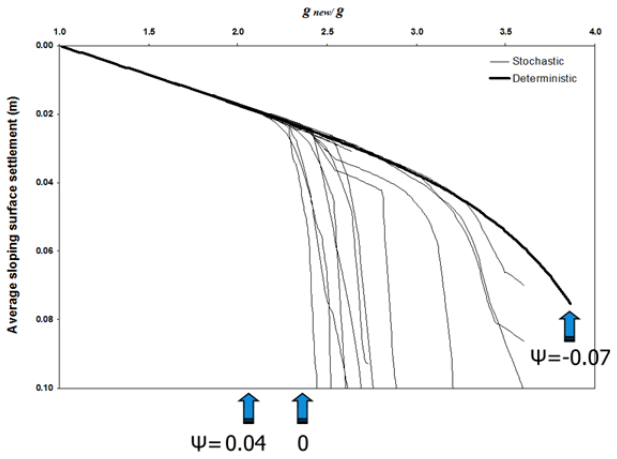


Figure 24: Influence of heterogeneity on undrained slope stability for μ_ψ = -0.07 [Bak01].

6 Conclusions

This chapter has shown how the random finite element method (RFEM) may be applied to practical problems, in order to gain insight into how soil heterogeneity influences material behaviour and geotechnical performance. It is apparent that, when taking account of heterogeneity in choosing characteristic soil property values, there are at least four components that should be considered. Firstly, the characteristic value is a function of both the point statistics and spatial correlation distances. Secondly, it is a function of the soil type. Thirdly, it is a function of the problem being analysed, including the problem geometry and loading and boundary conditions. Fourthly, it is a function of the extent of the labortaory and/or site investigation: that is, the more knowledge that is available about a site, the greater the degree of certainty about the likely range of structure responses.

References

[Arn01] Patrick Arnold, Gordon A. Fenton, Michael A. Hicks, Timo Schweckendiek & Brian Simpson (Eds.). *Modern geotechnical design codes of practice*. IOS Press, 2012.

[Arn02] Patrick Arnold & Michael A. Hicks. A stochastic approach to rainfall-induced slope failure. *Proc. 3rd Int. Symp. Safety and Risk, Munich, Germany, 107-115*, 2011.

[Bak01] Siamak Bakhtiari. *Stochastic finite element slope stability analysis*. PhD thesis, University of Manchester, 2011.

[Bee01] Ken Been & Michael G. Jefferies. A state parameter for sands, *Géotechnique, 35(2):99-112*, 1985.

[Cal01] Ed O. Calle. Probabilistic analysis of stability of earth slopes. *Proc. 11th Int. Conf. Soil Mechanics and Foundation Engineering, San Francisco, USA, 809-812*, 1985.

[CEN01] CEN. *Eurocode 7: Geotechnical design. Part 1: General rules*. EN 1997-1, European Committee for Standardization, 2004.

[DeG01] Don J. DeGroot & Gregory B. Baecher. Estimating autocovariance of in-situ soil properties, *J. Geotech. Eng., ASCE, 119:147-166*, 1993.

[Fen01] Gordon A. Fenton & D. Vaughan Griffiths. *Risk assessment in geotechnical engineering*. Wiley, 2008.

[Fen02] Gordon A. Fenton & Erik H. Vanmarcke. Simulation of random fields via local average subdivision, *J. Eng. Mech., ASCE, 116(8):1733-1749*, 1990.

[Fen03] Gordon A. Fenton & Erik H. Vanmarcke. Spatial variation in liquefaction risk, *Géotechnique, 48(6):819-831*, 1998.

[Hic01] Michael A. Hicks. Experience in calibrating the double-hardening constitutive model Monot, *Int. J. Num. Anal. Meth. Geomech., 27:1123-1151*, 2003.

[Hic02] Michael A. Hicks (Ed.). *Risk and variability in geotechnical engineering*. Thomas Telford, 2007.

[Hic03] Michael A. Hicks. AN explanation of characteristic values of soil properties in Eurocode 7. In: Arnold, Fenton, Hicks, Schweckendiek, Simpson (eds.), *Modern geotechnical design codes of practice, IOS Press, 36-45*, 2012.

[Hic04] Michael A. Hicks & Rached Boughrarou. Finite element analysis of the Nerlerk underwater berm failures, *Géotechnique, 48(2):169-185*,1998.

[Hic05] Michael A. Hicks & Jonathan D. Nuttall. Influence of soil heterogeneity on geotechnical performance and uncertainty: a stochastic view on EC7. *Proc. 10th Int. Probabilistic Workshop, Stuttgart, Germany, 215-227*, 2012.

[Hic06] Michael A. Hicks & Christakis Onisiphorou. Stochastic evaluation of static liquefaction in a predominantly dilative sand fill, *Géotechnique, 55(2):123-133*, 2005.

[Hic07] Michael A. Hicks & Kristinah Samy. Influence of heterogeneity on undrained clay slope stability, *Quarterly J. Engineering Geology and Hydrogeology, 35:41-49*, 2002.

[Hic08] Michael A. Hicks & Kristinah Samy. Reliability-based characteristic values: a stochastic approach to Eurocode 7, *Ground Engineering 35:30-34*, December, 2002.

[Hic09] Michael A. Hicks & Kristinah Samy. Stochastic evaluation of heterogeneous slope stability, *Italian Geotech. J., 38:54-66*, 2004.

[Hic10] Michael A. Hicks & William A. Spencer. Influence of heterogeneity on the reliability and failure of a long 3D slope, *Computers and Geotechnics, 37:948-955*, 2010.

[Hic11] Michael A. Hicks, Jonathan D. Nuttall & Jian Chen. Influence of heterogeneity on 3D slope reliability and failure consequence, *Computers and Geotechnics, 61:198-208*, 2014.

[Li01] Yajun Li & Michael A. Hicks. Comparative study of embankment reliability in three dimensions. *Proc. 8th Int. Conf. Num. Methods in Geotech. Eng., Delft, The Netherlands, 467-472*, 2014.

[Li02] Yajun Li, Michael A. Hicks & Jonathan D. Nuttall. Probabilistic analysis of a benchmark problem for slope stability in 3D. *Proc. 3rd Int. Symp. Computational Geomech., Krakow, Poland, 641-648*, 2013.

[Llo01] Marti Lloret-Cabot, Gordon A. Fenton & Michael A. Hicks. On the estimation of scale of fluctuation in geostatistics, *Georisk 8:129-140*, 2014.

[Llo02] Marti Lloret-Cabot, Michael A. Hicks & Abraham P. van den Eijnden. Investigation of the reduction in uncertainty due to soil variability when conditioning a random field using Kriging, *Géotechnique Letters 2:123-127*, 2012.

[Mol01] Fans Molenkamp. *Elasto-plastic double hardening model Monot*. LGM Report CO-218595, Delft Geotechnics, 1981.

[Nut01] Jonathan D. Nuttall. *Parallel implementation and application of the random finite element method*. PhD thesis, University of Manchester, 2011.

[Oni01] Christakis Onisiphorou. *Stochastic analysis of saturated soils using finite elements*. PhD thesis, University of Manchester, 2000.

[Pop01] Radu Popescu, Jean H. Prévost & George Deodatis. Effects of spatial variability on soil liquefaction: some design recommendations, *Géotechnique, 47(5):1019-1036*, 1997.

[Sam01] Kristinah Samy. *Stochastic analysis with finite elements in geotechnical engineering*. PhD thesis, University of Manchester, 2003.

[Smi01] Ian M. Smith, D. Vaughan Griffiths & Lee Margetts. *Programming the finite element method*. Wiley, 2013.

[Spe01] William A. Spencer. *Parallel stochastic and finite element modelling of clay slope stability in 3D*. PhD thesis, University of Manchester, 2007.

[Spe02] William A. Spencer & Michael A. Hicks. A 3D finite element study of slope reliability, *Proc. 10th Int. Symp. Num. Models Geomech., Rhodes, Greece, 539-543*, 2007.

[Tay01] Donald W. Taylor. Stability of earth slopes, *J. Boston Society of Civil Engineers, 24:197-246*, 1937.

[Van01] Erik H. Vanmarcke. Reliability of earth slopes, *J. Geotech. Eng. Div., ASCE, 103(11):1247-1265*, 1977.

[Van02] Erik H. Vanmarcke. *Random fields: analysis and synthesis*. The MIT Press, 1983.

[Wic01] Damika Wickremesinghe & Richard G. Campanella. Scale of fluctuation as a descriptor of soil variability, *Proc. Conf. Probabilistic Methods in Geotechnical Engineering, Canberra, Australia, 233-239*, 1993.

[Won01] Shiao Yun Wong. *Stochastic characterisation and reliability of saturated soils*. PhD thesis, University of Manchester, 2004.

Geotechnical back-analysis using a maximum likelihood approach

Alberto Ledesma

Universitat Politècnica de Catalunya (UPC-BarcelonaTech), Spain

Identifying parameters of a Geomechanical model from field measurements constitutes a typical "inverse problem" and is receiving increasing attention by the Geotechnical community. This text describes the basic concepts and procedures required to perform this inverse problem or back-analysis in a systematic manner. A Maximum Likelihood Approach is presented as a general framework and it is used to introduce the main topics involved. An objective function is defined, depending on the differences between the computed and measured variables. The minimum of that function corresponds to the parameters that best simulate the measurements. Some difficulties arise regarding uniqueness of the solution. The procedure is based on the sensitivity matrix computed as the derivatives of the measured variables with respect to the parameters. When some prior information on the parameters is available, the framework allows its inclusion in a consistent manner. Several examples based on the excavation of tunnel or underground caverns are used to illustrate the procedures described. Finally, some comments on other optimization techniques (i.e. genetic algorithms) and related topics, as model identification and optimal design of experiments are briefly addressed as well.

1 Introduction

In Geotechnical Engineering, the determination of soil or rock parameters has been usually performed by means of laboratory or field tests. Also, it has been quite common to use back-analysis to obtain material parameters in the context of failures and forensic geotechnics [Puz10]. In the last two decades, simultaneously to the development of numerical methods, new techniques to determine material parameters from field measurements have been proposed. These techniques were first used in Geophysics and in Groundwater Hydrology, where it is not possible to characterize properly the material in the laboratory.

In a typical problem, for a particular geometry, we estimate displacements and pore water pressures generated by a change in stresses and/or hydraulic or mechanical boundary conditions. To do that, we assume a constitutive law for the materials and the corresponding parameters. In this manner we are solving what is called "the direct problem". Sometimes, however, we may know some displacements and/or pore water pressures at some specifics points of the geometry and we want to know the parameters of the model or the model itself. Then, we are solving the "inverse problem". In a broad perspective the "inverse problem" may refer also to identifying the geometry or the boundary conditions.

Nowadays, the most common inverse problem in Geomechanics is the identification of parameters of a fixed model from measured displacements. This procedure seems to be fully adapted to Geotechnical Engineering practice, where sometimes decisions are taken based on field measurements, as proposed by Terzaghi for the "observational method" [Ter67]. By using field instrumentation measurements to estimate geotechnical parameters, it is possible to take into account the large scale structure of the soil or rock, which is outside the possibilities of the other procedures of parameter identification.

The identification of parameters results in an optimization problem from a mathematical point of view. In recent times, the development of numerical methods and optimization techniques has allowed a more systematic and rational approach to this problem. Thus the use of the "observational method" combined with this parameter identification techniques could be used in practical applications during the construction stage. The case of a tunnel excavation has been adopted here as a typical example, as measurements are usually performed in a continuous manner. Therefore material parameters can be identified and compared with the initial assumptions and decisions could be taken for further excavation stages.

It is important to note that the process of parameter estimation presented here is carried out in the context of a specified fixed model that includes geometry, boundary conditions and constitutive laws for the materials. Therefore, the differences between the measurements and the predictions of the model are assumed to be due to a measurement error, as the model is considered correct. The topic of selecting the best model or the identification of the model itself is not treated here, although it will be briefly commented. Also, the decision about what and where we should take measurements could be assisted by the information provided by these identification techniques, as they provide an insight into the model structure.

In this text, the Maximum Likelihood Framework is used to present the main concepts related to back-analysis. Also some examples based on tunnel excavation problems are used for illustration purposes. Finally, a new hybrid method combing genetic algorithms and a gradient-based algorithm is presented, which is particularly useful for problems with a large number of parameters being identified.

2 Maximum Likelihood Approach

There are several approaches to cope with the identification of parameters in Geomechanics. Many significant contributions in this topic have used a deterministic point of view, but there are important advantages in adopting a probabilistic framework ([Civ83], [Gen88]). Here, a probabilistic approach based on the maximum likelihood concept is presented. There are other alternatives, but in most cases they result in the same mathematical formulation, being different the conceptual framework only.

2.1 Basic formulation

Let us assume that a deterministic model, M, relates some unknown parameters, **p**, and some variables, **x**, that is, **x** = M(**p**). The measurements are represented by **x***. Then, the differences between measurements and predictions of the model (**x***-**x**) are considered as an error that can be defined in a probabilistic manner. The best estimation of the parameters is assumed to be found by maximizing the likelihood, *L*, of a hypothesis, **p**, given a set of error measurements, (**x***-**x**). The likelihood of a hypothesis is proportional to the conditional probability of **x*** given a set of parameters **p** ([Edw72], [Tar87]):

$$L = kf(\mathbf{x}^*/\mathbf{p}) \tag{1}$$

where *k* is a proportionality constant. This formulation has conceptual advantages, in particular: it does not require reproducing the true system exactly and the differences between field measurements and model predictions are due to error measurement [Led96a]. Therefore, the probability of measuring **x*** given a set of parameters **p**, is the probability of reproducing the error measurements (**x***-**x**). Assuming that probability distribution as multivariate Gaussian, it is possible to write:

$$P(\mathbf{x}^* - \mathbf{x}) = \frac{1}{\sqrt{(2\pi)^m\,|\mathbf{C_x}|}}\, exp\left[-\frac{1}{2}(\mathbf{x}^* - \mathbf{x})^t(\mathbf{C_x})^{-1}(\mathbf{x}^* - \mathbf{x})\right] \tag{2}$$

where $\mathbf{C_x}$ is the measurements covariance matrix representing the structure of the error measurements, *m* is the number of measurements and $(\ )^t$ represents a transposed matrix. The likelihood is now proportional to the value expressed by (2). Maximizing *L* is equivalent to minimize the "support function": $S = -2\ \ln L$ and it follows:

$$S = (\mathbf{x}^* - \mathbf{x})^t(\mathbf{C_x})^{-1}(\mathbf{x}^* - \mathbf{x}) + \ln|\mathbf{C_x}| + m\ln(2\pi) - 2\ln k \tag{3}$$

If the error structure of the measurements is fixed, only the first term in (3) has to be minimized, and the final expression is defined as the "objective function":

$$J = (\mathbf{x}^* - \mathbf{x})^t(\mathbf{C_x})^{-1}(\mathbf{x}^* - \mathbf{x}) \tag{4}$$

If the error structure is not fixed, the problem becomes more complex, although in some cases it is possible to incorporate the error structure to the set of parameters to be identified [Led96b]. Equation (4) becomes simpler when the measurements are independent and their errors have a Gaussian distribution of probability with the same variance, σ_x^2, because $\mathbf{C_x} = \sigma_x{}^2\mathbf{I}$, being $\mathbf{I}$ the identity matrix. The objective function to be minimized results in

$$J = (\mathbf{x}^* - \mathbf{x})^t(\mathbf{x}^* - \mathbf{x}) \tag{5}$$

Expression (5) represents in fact a least squares criterion and the parameters that best characterize the model are those that minimize the square differences between the measured and the computed variables. Now the value of σ_x^2 is not required to obtain the minimum of (5). This approach has become very popular in Geotechnical back-analysis as field measurements can be considered independent and following a Gaussian distribution of probability in many practical situations. In some cases, however, measurements are not independent and the corresponding covariance matrix should be used (i.e. when an inclinometer is used, because horizontal movements are obtained by adding previous incremental movements) [Led96a]. Also, if there are different types of measurements, each type should have its own covariance matrix.

2.2 Numerical implementation

The minimum of the objective function has to be obtained using a suitable numerical procedure. As the Finite Element Method is the most popular procedure to solve the "direct problem", it is convenient to combine the algorithm to minimize the objective function, *J*, and the Finite Element code. In general, minimization or optimization algorithms can be classified in two groups: those that need to compute the gradient of *J*, and those that evaluate *J* only. It may be expected that algorithms using the derivative of the objective function are more robust and powerful, although more difficult to compute, than algorithms using only the values of the function. In the last 30 years, examples of both types of algorithms applied to Geomechanics have been published. Downhill simplex, a typical procedure for unconstrained optimization that does not require computing derivatives, was used already by [Gio80]. Genetic algorithms have proven to be convenient in this context and do not require computing the gradient either [Lev09].

Here, the Gaus-Newton method based on the computation of the gradient of *J* is presented. It has good convergence properties and the derivatives obtained are also useful in providing information on the reliability of the parameters identified. The procedure is iterative and from a set of parameters at iteration *k*, the new set of parameters is:

$$\mathbf{p}_{k+1} = \mathbf{p}_k + \Delta\mathbf{p}_k \quad , \quad J(\mathbf{p}_{k+1}) \leq J(\mathbf{p}_k) \tag{6}$$

The Gaus-Newton method [Fle81] is based on an expansion of the objective function into a Taylor series, and gives the value of the advance vector in the parameters space, $\Delta\mathbf{p}$ in iteration *k*, according to:

$$\Delta\mathbf{p} = (\mathbf{A}^t\mathbf{C}_x^{-1}\mathbf{A})^{-1}\mathbf{A}^t\mathbf{C}_x^{-1}\Delta\mathbf{x} \tag{7}$$

where $\Delta\mathbf{x}=(\mathbf{x}^*-\mathbf{x})$ and the matrix $\mathbf{A}=\partial\mathbf{x}/\partial\mathbf{p}$ is called the sensitivity matrix. If the number of measures is *m* and the number of parameters is *n*, the size of this matrix is m×n. When there is any convergence problem, an improvement of the algorithm proposed by Levenberg and Marquardt can be used [Mar63], in which equation (7) is changed into:

$$\Delta\mathbf{p} = (\mathbf{A}^t\mathbf{C}_x^{-1}\mathbf{A} + \mu\mathbf{I})^{-1}\mathbf{A}^t\mathbf{C}_x^{-1}\Delta\mathbf{x} \tag{8}$$

where μ is an arbitrary real number. If the value of *J* becomes smaller in the next iteration, μ is decreased, reaching 0 or a very small value at the minimum. However, if *J* increases then μ is also increased and the increment of parameters obtained by (8) tends towards the gradient of the objective function. The initial value of μ and the manner it is increased or decreased has to be defined in advance and sometimes trial identifications are required.

2.3. Coupling to the finite element method

The procedure outlined above should be associated with a numerical model relating measurements and parameters. The finite element Method is appropriate for this and it is possible to combine the Gauss-Newton technique with the Finite Element formulation. A simple case that may be considered first refers to some nodal displacements being measured and a linear isotropic elastic model characterized by Young's modulus and Poisson's ratio. The finite element method gives a linear system of equations:

$$\mathbf{Kx} = \mathbf{f} \quad , \quad \mathbf{K} = \int_V \mathbf{B}^t\boldsymbol{D}\boldsymbol{B}dV \quad , \quad \boldsymbol{f} = \int_S \mathbf{N}^t\boldsymbol{\sigma}dS \tag{9}$$

where **K** is the global stiffness matrix, **x** is the vector of nodal displacements, **f** is the nodal forces vector, **B** is the geometry matrix relating strains and nodal displacements, **D** contains the constitutive law as a relationship between stresses and strains and **N** is the shape function matrix. The vector $\boldsymbol{\sigma}$, for an excavation problem, is the vector of stresses acting on the excavation boundary, *S*.

Expressions (6) and (8) represent the iteration procedure to minimize the objective function defined by (4) or (5). Note that in (8) it is necessary to compute the sensitivity matrix **A**. For this particular case, it is possible to obtain an "exact" evaluation of **A** in the context of the finite element approximation. Deriving the first expression in (9) with respect to the parameters and rearranging, we obtain:

$$\frac{\partial\mathbf{x}}{\partial\mathbf{p}} = \mathbf{K}^{-1}\left(\frac{\partial\boldsymbol{f}}{\partial\mathbf{p}} - \frac{\partial\mathbf{K}}{\partial\mathbf{p}}\mathbf{x}\right) \tag{10}$$

If the parameters to be identified are Young modulus and Poisson's ratio, then $\partial\mathbf{f}/\partial\mathbf{p}=0$, and

$$\frac{\partial \mathbf{x}}{\partial \mathbf{p}} = -\mathbf{K}^{-1}\left(\frac{\partial \mathbf{K}}{\partial \mathbf{p}}\mathbf{x}\right) \quad , \quad \frac{\partial \boldsymbol{K}}{\partial \mathbf{p}} = \int_V \mathbf{B}^t \frac{\partial \mathbf{D}}{\partial \mathbf{p}} \mathbf{B} dV \tag{11}$$

Hence computing $\partial\mathbf{K}/\partial\mathbf{p}$ is equivalent to find the stiffness matrix substituting the $\partial\mathbf{D}/\partial\mathbf{p}$ matrix for $\mathbf{D}$ which allows an easy implementation in a finite element code. In particular, if Poisson's ratio is assumed known and only Young's modulus, E, is identified, $\partial\mathbf{D}/\partial E$ is constant and has to be calculated only once.
When the identification of the earth pressure coefficient at rest, K_o, is required, equation (10), for a linear elastic material, results in

$$\frac{\partial \mathbf{x}}{\partial K_o} = \mathbf{K}^{-1}\frac{\partial \mathbf{f}}{\partial K_o} \tag{12}$$

since the stiffness matrix is independent of K_o in that case. The excavation is simulated in one phase, by applying on the excavated boundary the opposite normal forces to the initial stresses. So, from the last equation in (9) we obtain:

$$\frac{\partial \mathbf{f}}{\partial K_o} = \int_S \mathbf{N}^t \frac{\partial \boldsymbol{\sigma_o}}{\partial K_o} dS = \int_S \mathbf{N}^t \boldsymbol{\sigma_o^*} dS \tag{13}$$

where for typical plane strain problems, $\boldsymbol{\sigma_o}=(K_o\sigma_y^o, \sigma_y^o, 0)$, σ_y^o is the initial vertical stress at depth "y" and $\boldsymbol{\sigma_o^*} = (\sigma_y^o, 0, 0)$. Note that calculating expression (12) becomes simple as the finite element routines used to find nodal forces due to excavation can be employed to compute $\partial\mathbf{f}/\partial K_o$ by changing only the initial stress field from $\boldsymbol{\sigma_o}$ to $\boldsymbol{\sigma_o}^*$.
This procedure is expected to reduce the numerical errors when calculating the sensitivity matrix **A**. However, in a general problem involving nonlinear constitutive models a finite difference technique is the common procedure to estimate the derivatives of some nodal variables with respect to the parameters. Using a central difference scheme, two direct problems have to be solved to compute one derivative as:

$$\left.\frac{\partial \mathbf{x}}{\partial p_i}\right|_{\mathbf{p}=\mathbf{p}_k} = \left.\frac{\mathbf{x}(p_i+\Delta p_i)-\mathbf{x}\left((p_i-\Delta p_i)\right)}{2\Delta p_i}\right|_{\mathbf{p}=\mathbf{p}_k} \tag{14}$$

where Δp_i is the value used to increment parameter "i" while keeping the rest of the parameters constant. Defining this increment requires a trial and error procedure in advance in order to avoid additional numerical errors. The procedure seems to be less robust than the one presented above in equations (9) to (13), where there is a direct use of the finite element approximation.

2.4. Reliability of the estimation

One of the advantages of using gradient-based procedures to minimize the objective function is that the derivatives provide with information on the reliability of the estimation. It can be shown [Led96] that a lower bound of the covariance matrix of the identified parameters, $\mathbf{C_p}$ can be obtained as:

$$\mathbf{C_p} = \left(\mathbf{A}^t \mathbf{C_x}^{-1} \mathbf{A}\right)^{-1} \tag{15}$$

This matrix agrees with the inverse of the Fisher information matrix, a classical concept in optimization and parameter identification [Bur75], [Wal90]. If a particular measurement is not sensitive to a particular parameter, the corresponding derivative of the sensitivity matrix will tend to zero and the inverse matrix in (15) will provide a large variance for that particular parameter.

It follows from the above expressions that the sensitivity matrix includes a lot of information about the structure of the model. Note that typically we have more measurements, m, than parameters to identify, n, so therefore, $\mathbf{A}$ is a rectangular matrix $m \times n$. If the number of measurements coincides with the number of parameters, there is a unique solution and the problem becomes deterministic.

The value of the objective function at the minimum, J_{min}, indicates the global error of the problem. When all measurements are of the same type, it is possible to estimate the standard deviation of the measurements, σ_x , by means of [Wig72]:

$$\sigma_x = \sqrt{\frac{J_{min}}{m-n}} \tag{16}$$

Note that this value represents the error assigned to the measurements by the procedure because the model is assumed to be "perfect". In practice, measurements may have less error than that and discrepancy between measurements and the computed values may be due to the model adopted in the analysis not representing the real behaviour of the material.

3. A synthetic example involving a tunnel excavation

To show the capabilities of the formulation and to illustrate some of the difficulties related to back-analysis, a synthetic and simple example involving the excavation of a tunnel is presented. Two cases are considered, depending on the constitutive law used.

3.1. Case 1. Linear elastic model.

A case presented in [Led96] is described here. Only one material is considered and it is assumed linear elastic, homogeneous and isotropic with a Poisson's ratio of 0.49

(simulating undrained conditions) and specific weight of 20 kN/m^3. The parameters to be identified are the Young's modulus and the K_o coefficient defined using total stresses. Figure 1 presents the finite element mesh used in this case including 12 nodal points where displacements are supposed to be measured. Horizontal displacements in points 1 to 7 represent inclinometric measurements, whereas vertical movements in points 8 to 12 represent displacements obtained from an extensometer. Excavation is made in one step and only half of the geometry is considered due to symmetry.

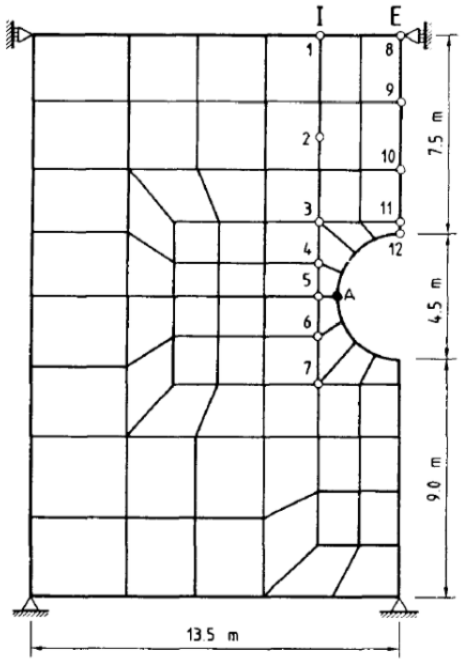


Figure 1. Finite element mesh and measurement points (after [Led96]).

The case is considered "synthetic" in the sense that measurements are in fact obtained by computing the direct problem for the parameters E=10 MPa and K_o=1. The values of these theoretical measurements are presented in Table 1. There is not error measurement and the minimum of the objective function should be zero. Because of that, the covariance matrix is assumed to be the identity matrix.

Table 1. Computed displacements used as measurements in the synthetic example (after [Led96]).

Horizontal movement	**1**	**2**	**3**	**4**	**5**	**6**	**7**
	0.316	0.556	2.038	3.550	4.550	3.920	2.427
Vertical movement	**8**	**9**	**10**	**11**	**12**	Values in cm	
	-3.598	-3.905	-4.935	-6.327	-7.048		

When only two parameters are identified, it is possible to depict the objective function and to observe the evolution of the iterations. Figure 2 presents contours of the objective function in terms of $E - K_o$. The values E=0.5 MPa and K_o=0.5 were used as initial parameters in the iterative procedure. Two paths have been depicted in the figure, corresponding to a Gauss-Newton algorithm and to a Levenberg-Marquardt algorithm. It can be seen that in a few iterations the minimum corresponding to E = 10 MPa and K_o=1 is reached. The objective function has a paraboloid-type shape, which is very convenient for the Gauss-Newton algorithm and, in general, suggests that the problem is well-posed and the solution is unique.

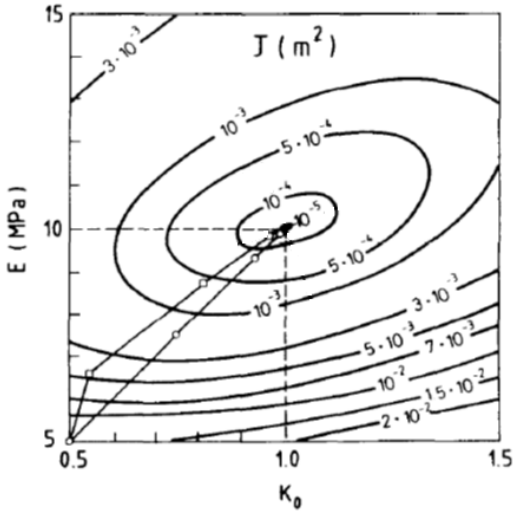


Figure 2. Contours of the objective function for the simpler synthetic case, indicating the paths followed by the Gauss-Newton (circles) and Marquardt (squares) algorithms (after [Led96]).

Let us consider now that the number of measurements is increased. The procedure should indicate to some extent that there is more information available and thus the error of the parameters identified should be smaller. In this synthetic case, as there is not error in the measurements, we can check the value of expression (15) using the identity matrix as covariance matrix. Four analyses have been carried out considering an increasing amount of measurements:

a) Case with only two measurements, vertical displacement at point 12 and horizontal displacement at point A in Figure 1.
b) Case with 12 measurements as described above.
c) Case with 24 measurements, 15 horizontal from nodes located on the vertical line I (figure 1) and 9 vertical located on the vertical line E (figure 1).
d) Case with 55 measurements. 24 of them are the same as in previous case and the rest are horizontal and vertical movements from points distributed on the excavation boundary.

For these cases the value of expression (15) has been computed at the minimum. The parameters at that minimum correspond to E = 10 MPa and K_o = 1 in all cases, but the shape of the objective function is different and thus the variance of the parameters computed. Figure 3 shows the standard deviations of the parameters (proportional to the measurements's standard deviation) for the four cases considered. As expected, increasing the number of measurements improves the quality of the identification, reducing the variance of the parameters identified.

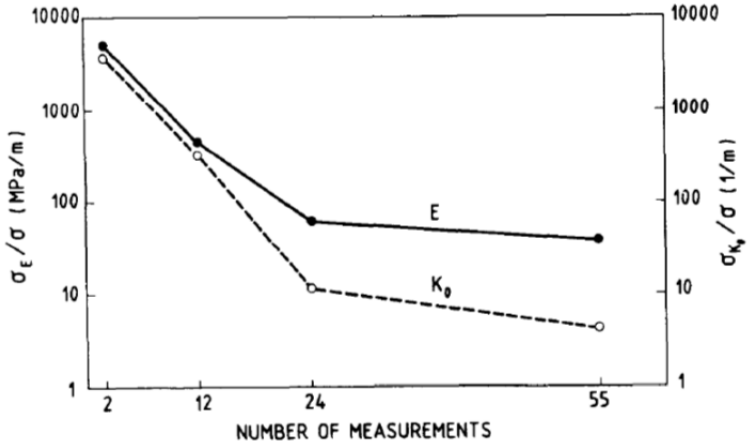


Figure 3. Standard deviations of E and K_o when different number of measurements are used in the identification process (after [Led96]).

3.2. Case 2. Nonlinear elastic model.

An example presented in [Led91] is shown here, in order to explain a case with a nonlinear elastic constitutive law and to comment some aspects regarding the uniqueness of the solution. Consider the same geometry indicated in previous example and the same measurement points (Figure 1). Now the soil is assumed to behave according to a hyperbolic model (nonlinear elastic), based on the proposal by Duncan and Chang [Dun70]. The model can be formulated in terms of secant stiffness or a secant constitutive matrix, $\mathbf{D_S}$. For plane strain problems and considering undrained conditions that matrix can be expressed as:

$$\mathbf{D_S} = \frac{E_i\left(1-R\frac{J_T}{c_u}\right)}{(1+\nu)(1-2\nu)}\begin{bmatrix}(1-\nu) & \nu & 0\\ \nu & (1-\nu) & 0\\ 0 & 0 & \frac{(1-2\nu)}{2}\end{bmatrix} \tag{17}$$

where E_i is the initial Young modulus, ν the Poisson's ratio, c_u the undrained shear strength, R is a parameter controlling whether c_u is reached asymptotically or not and J_T is the second invariant of the deviator stress tensor:

$$J_T{}^2 = \frac{1}{2} tr\mathbf{S}^2 \quad ; \quad S_{ij} = \sigma_{ij} - p\delta_{ij} \tag{18}$$

where σ_{ij} is the stress Cauchy tensor, δ_{ij} is the Kronecker delta, p is the mean stress and tr means trace. In this case, a value of R=1 was adopted and a total stress approach was used due to the undrained conditions. The advantage of a nonlinear elastic model in this context is that it is still possible to compute the sensitivity matrix in an "exact" manner within the finite element framework. Details of that computation are presented in [Led91].
Let us consider two parameters to identify, E_i and c_u, and the set of 12 displacements measured at the points indicated in Figure 1. The "measured" displacements will correspond to the values obtained by computing the direct problem for the following cases:

a) E_i = 10 MPa , c_u = 0.1 MPa
b) E_i = 30 MPa , c_u = 0.3 MPa
c) E_i = 50 MPa , c_u = 0.5 MPa
d) E_i = 100 MPa , c_u = 1.0 MPa

The ratio E_i/c_u has been kept constant in all cases. Measurements generated have been assumed to be independent and with the same variance, so the covariance matrix does not affect the identification process.
According to [Dav80], a value of $c_u \approx 0.06$ MPa would imply collapse for the particular geometry adopted. Therefore, case a) can be considered close to failure conditions, whereas case d) is far from collapse and the elastic behaviour will dominate.

Contours of equal value of the objective function for case a), close to failure, are depicted in Figure 4(left). The shape of those contours suggest that the problem is highly nonlinear: there is a sort of "valley" in which there are several combinations of parameters that provide almost the same objective function. This difficulty regarding uniqueness of the solution is quite typical in optimization problems. The Figure shows also two iterative paths followed by the algorithm, one of them failing in obtaining the minimum. The shape of the objective function is dominated by valleys following the direction of the E_i axis and this is because close to failure, that parameter becomes more difficult to estimate.

Case d), far from failure, is also presented in Figure 4 (right). Note that contours of the objective function are almost parallel to c_u axis, thus making difficult the identification of that parameter. In fact, the algorithm was not able to obtain a good value of c_u. That could be expected, as we are trying to identify a parameter controlling failure in a problem mainly elastic. The formulation of the hyperbolic model involves both parameters simultaneously and this is why the contours are not totally parallel to one of the axis. However, the shape of the objective function can be even worse (regarding the difficulty for minimization) if the constitutive model separates

elastic and plastic behaviour. This example suggests that in the context of Geomechanics, the identification of parameters by using any optimization procedure may require some previous experience regarding which parameters are important and which are not relevant in the problem. Mathematically ill-posed problems may be just a consequence of improper initial assumptions, i.e. we cannot identify plastic parameters if the measured displacements are in the elastic range.

The identification process gives information on the quality of the parameters identified, through expression (15). Table 2 presents the ratio of variances of the parameters identified, that is, *var* E_i*/var* c_u, for the cases considered. It can be seen that in case a), close to failure, that ratio is quite large, whereas in case d), far from failure, the ratio becomes small.

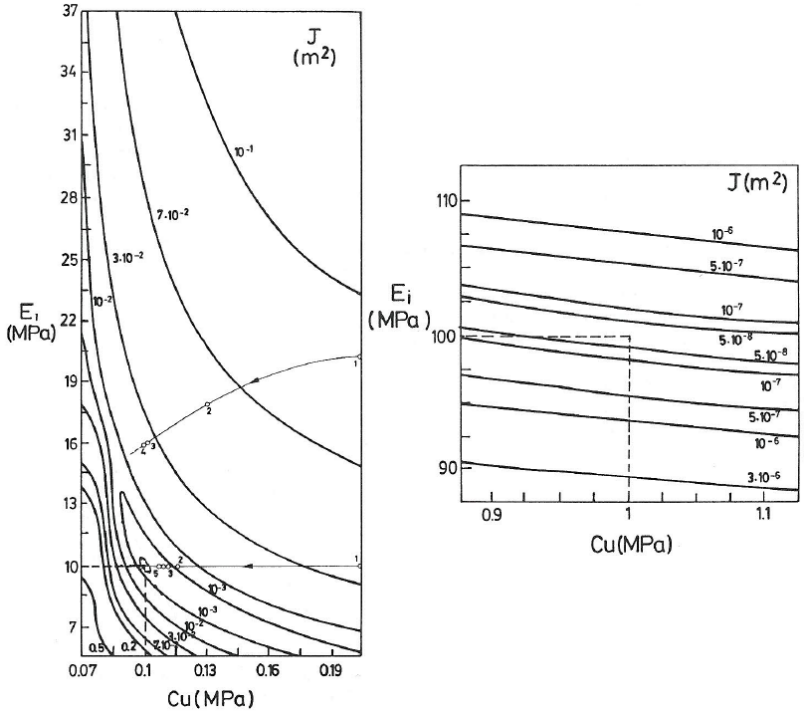


Figure 4. Left: Objective function for case a), close to failure. Right: Objective function for case d), far from failure, (after [Led91]).

Table 2. Ratio of variances of the parameters identified, *var* E_i*/var* c_u.

Case	a)	b)	c)	d)
var E_i*/var* c_u	62405	1833	537	111

4. Including Prior Information

4.1. Extension of the formulation

It is quite common to have some prior information on the parameters before proceeding with the identification process, typically from laboratory or field tests. The maximum likelihood approach can be extended to incorporate that information in the identification process in a consistent manner. Now it is assumed that prior information of the parameters $< \mathbf{p} >$ has a multivariate Gaussian probability distribution. Therefore, it is possible to write:

$$P(< \mathbf{p} > - \mathbf{p}) = \frac{1}{\sqrt{(2\pi)^n \, |\mathbf{C_p}^o|}} exp\left[-\frac{1}{2}(< \mathbf{p} > - \mathbf{p})^t \left(\mathbf{C_p}^o\right)^{-1} (< \mathbf{p} > - \mathbf{p})\right] \quad (19)$$

where $\mathbf{C_p}^{\mathbf{o}}$ is the "a priori" parameter covariance matrix, based on the available prior information and n is the number of parameters. Assuming that the measurements $\mathbf{x}^*$ and the parameters given by the prior information $< \mathbf{p} >$ are independent, expression (1) is modified to

$$L = kP(\mathbf{x}^* - \mathbf{x})P(< \mathbf{p} > - \ \mathbf{p}) \quad (20)$$

Then the support function becomes

$$S = (\mathbf{x}^* - \mathbf{x})^t (\mathbf{C_x})^{-1} (\mathbf{x}^* - \mathbf{x}) + (< \mathbf{p} > - \mathbf{p})^t \left(\mathbf{C_p^o}\right)^{-1} (< \mathbf{p} > - \mathbf{p}) + \ \ln|\mathbf{C_x}| +$$

$$+ \ \ln\left|\mathbf{C_p^o}\right| + m\ln(2\pi) + n\ln(2\pi) - 2\ln k \quad (21)$$

If the error structure of measurements and parameters are considered fixed, only the first two terms must be used in the minimization process, the rest being constant. Those two terms define the objective function. As stated before, any optimization algorithm can be used. In particular, the Gauss-Newton and Marquardt algorithms defined above can be extended to take into account the new second term in (21), [Gen88]. Then equation (8), used to iterate in the parameters space, modifies to

$$\Delta\mathbf{p} = \Delta\mathbf{p^o} + \left[\mathbf{A}^t \mathbf{C_x^{-1}} \mathbf{A} + \left(\mathbf{C_p^o}\right)^{-1} + \mu\mathbf{I}\right]^{-1} \mathbf{A}^t \mathbf{C_x^{-1}} (\Delta\mathbf{x} - \mathbf{A}\Delta\mathbf{p^o}) \quad (22)$$

where $\Delta\mathbf{p^o} = < \mathbf{p} > - \mathbf{p}$.

4.2 Example involving the staged excavation of a cavern in rock

To illustrate the use of prior information, an example presented in [Gen88] is summarized here. It involves the staged excavation of a powerhouse cavern in the Spanish Pyrenées. A cross section near the central part of the cavern was selected for the analysis so that plane strain conditions could be adopted. The excavation was performed in nine successive steps, but in order to have significant measurement val-

ues, they were grouped in 3 stages. Figure 5 presents the section considered, showing measurements locations and excavation phases. Displacements were obtained by means of extensometers and convergence measurements. In both cases they are relative displacements measured in one particular direction between two points. Therefore, measurements $\mathbf{x}^*$ does not correspond to nodal displacements, and a transformation has to be carried out for each measurement, x_i^*, according to:

$$x_i^* = \mathbf{L}_i \mathbf{T}_i \mathbf{N}_i \mathbf{u}_i \tag{23}$$

where $\mathbf{u}_i$ is the vector of nodal displacements of the elements containing measurements points, $\mathbf{N}_i$ is the shape function vector for the same elements, $\mathbf{T}_i$ is the rotation matrix required to transform the components of displacements in the global coordinate system to the measurement direction, and $\mathbf{L}_i$ is the matrix containing the linear combination of displacements corresponding to the type of measurement. When the measurement is a relative displacement between two points, $\mathbf{L}_i = (1,-1)$. Finally the sensitivity matrix can be computed as

$$\mathbf{A} = \frac{\partial \mathbf{x}}{\partial \mathbf{p}} = \mathbf{LTN} \frac{\partial \mathbf{u}}{\partial \mathbf{p}} \tag{24}$$

A total of 36 different movements were considered, of which 8 correspond to the first phase, 7 to the second and 21 to the third. Note (figure 5) that some instruments were installed when stage 2 was finished, so only part of the records are available for the 3 phases.

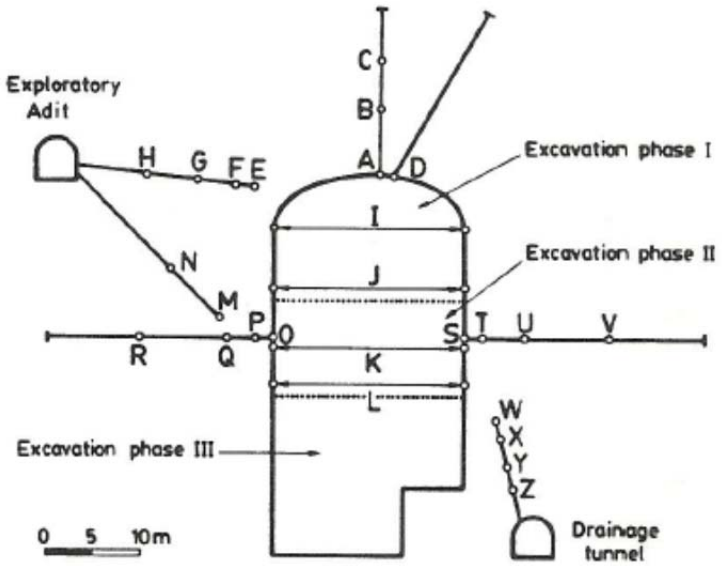


Figure 5. Underground cavern: geometry of the cross section considered in the analyses, excavation phases and measurement points (I,J,K and L are convergences).

The rock in the area is schist which shows no oriented texture due to the high degree of metamorphism to which was subjected. No significant anisotropy was detected in the field data, so the material was assumed linear elastic and isotropic. Poisson's ratio was assumed to be 0.28, based on laboratory experiments on rock samples.
One of the major uncertainties of the project was the initial stress state in the rock massif. In situ measurements prior to the excavation suggested that initial stresses were oriented with the cavern axis, being the vertical stress equal to the overburden pressure. Therefore it was decided to identify *Ko* (ratio between horizontal and vertical stresses, working in total stresses) from field measurements. The other one was Young's modulus, *E*, as a typical parameter of the elastic model. Only two parameters were considered in the identification problem, because it is possible to plot the objective function and to analyze the structure of the problem in a visual manner.

First, let us consider the case where only displacements due to the excavation are available. There is not prior information, and the measurements can be assumed independent with the same variance, so $\mathbf{C_x} = \sigma_x{}^2\mathbf{I}$. The objective function becomes the simple least squares criterion indicated in equation (5). To illustrate the structure of the problem, the objective function has been depicted in figure 6 (left). Note that there is a sort of valley in which different combinations of *E* and *Ko* give similar result in terms of squared error. Figure 6 (right) includes a section of the objective function along the valley, to highlight the difficulty in finding a global minimum.

The identification process yielded the values: E=0.39·10^4 MPa and *Ko*=1.24, with a minimum of J=2.89·10^{-4} m^2. Expression (16) in this case results in an estimation of the standard deviation of σ_x=2.9 mm. A comparison between computed and observed displacements is presented in Figure 7. Note that the parameters found and the model considered achieve a good general approximation to the observed displacements.

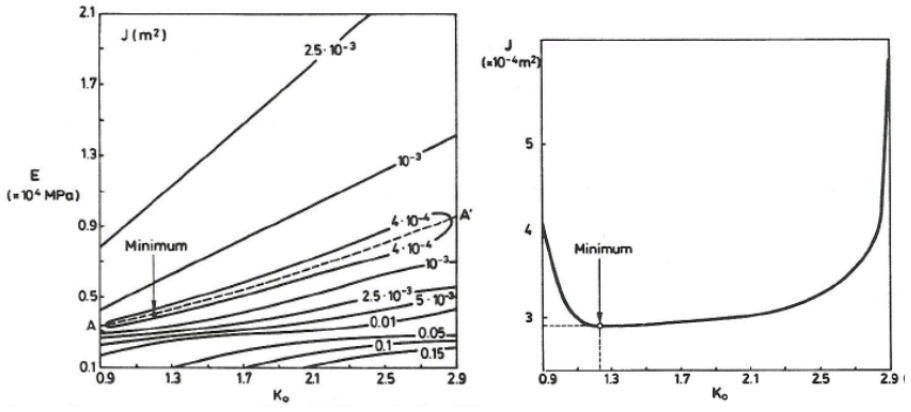


Figure 6. Left: Contours of the objective function J. Right: Values of the objective function along the valley (A-A').

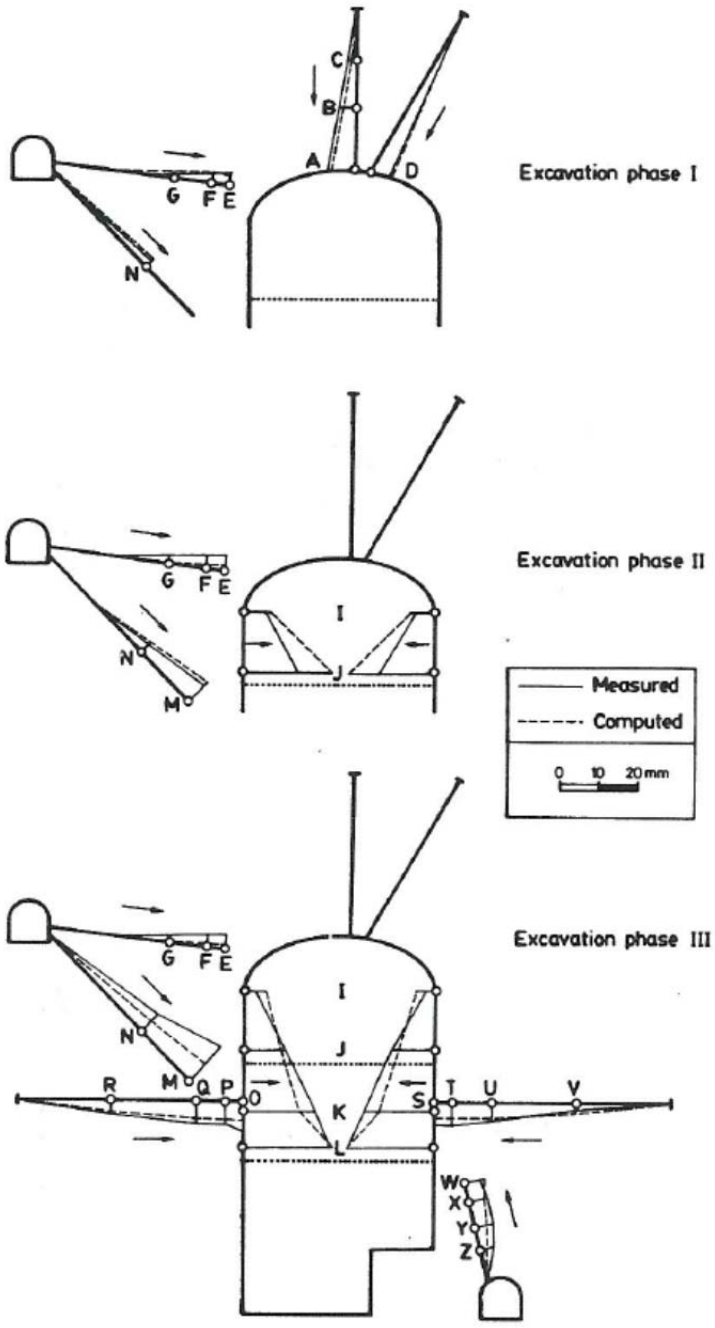


Figure 7. Measured and computed displacements with the parameters identified without prior information.

Despite the reasonable agreement between computed and measured displacements, the parameters identified were not consistent with the information obtained during the design of the cavern. Indeed most of the previous "in situ" tests suggested higher values for *E* and *Ko.* The Young's modulus of the rock was determined during the design phase using flat jack tests and dilatometer tests. Different orientations were considered, but the stiffness was isotropic.

A mean value from all measurements was $< E > = 1.5 \cdot 10^4$ MPa. The "in situ" stresses were measured using both flat jack tests and solid inclusion triaxial cells. Vertical stresses were always closed to overburden pressure, as expected. However, flat jack tests gave values of K_o around 1 and solid inclusion cells provide values of K_o in the range 2 to 3. A mean value of $< K_o > = 2$ was finally adopted.

The first two terms in (21) define the objective function in this case. Considering the prior information on the Young modulus, $<E>$, and on the initial stress ratio, $< Ko>$, the objective function can be written as

$$J = (\mathbf{x}^* - \mathbf{x})^t (\mathbf{x}^* - \mathbf{x}) + \left(\frac{\sigma_x^2}{\sigma_E^2}\right)(< E > - E)^2 + \left(\frac{\sigma_x^2}{\sigma_{K_o}^2}\right)(< K_o > - K_o)^2 \qquad (25)$$

where σ_x^2 is the variance of the measurements, σ_E^2 is the variance of the prior information on Young's modulus and σ_{Ko}^2 is the variance of the prior information on K_o coefficient. Note that the value of $(\sigma_E^2 / \sigma_x^2)$ to be used in the analysis depends not only on the scatter of the results of the "in situ" tests used to determine *E*, but, also, on the weight that is to be assigned to the prior information. In this case, a value of $(\sigma_E^2 / \sigma_x^2) = 10^4$ MPa^2/m^2 was adopted. If $\sigma_x \approx 3$ mm (as estimated in the analysis without prior information), the range of likely *E* values will be:

$$E = < E > \pm 2\sigma_E = 1.5 \cdot 10^4 \pm 0.6 \cdot 10^4 \; MPa \qquad (26)$$

Regarding K_o coefficient, a significant scatter was observed in the field tests. Therefore, the value $(\sigma_{Ko}^2 / \sigma_x^2) = 4 \cdot 10^4$ m^{-2} was used. Assuming again the value of $\sigma_x \approx 3$ mm, this gives a likely range of K_o:

$$K_o = < K_o > \pm 2\sigma_x = 2 \pm 1.2 \qquad (27)$$

With these conditions the minimum of the objective function (25) was found for the point: $E = 0.77 \cdot 10^4$ MPa and $K_o = 2.36$. The value of the minimum was $3.24 \cdot 10^{-4}$ m^2 which corresponds to a standard deviation of 3.1 mm by using expression (16). This value of 3.1 mm is very close to the standard deviation obtained when no prior information was considered.

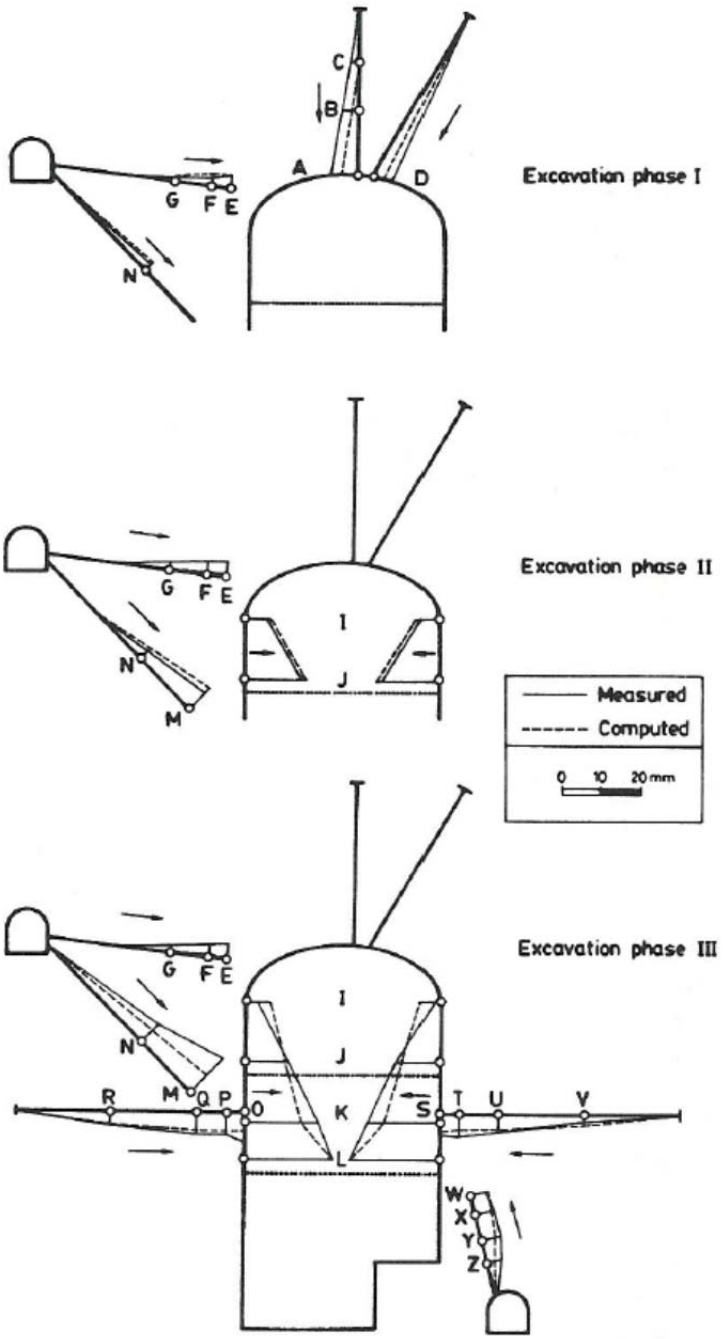


Figure 8. Measured and computed displacements using parameters identified with prior information.

Figure 8 presents the comparison between measured and computed displacements using this new set of parameters identified. The comparison is as reasonable as the one presented in Figure 7. It can be stated, however, that now the solution incorporates in a global manner all the information available: a priori field tests from the design phase and measurements obtained during the construction stage. Note that the identified K_o lies in the likely range, according to (27). However, E, lies outside the range expected in (26). There may be several reasons explaining that discrepancy.

The most important one is related to the assumptions of the model and the actual rock behaviour. Note that field measurements during construction take larger times compared to "in situ" tests and that may be the reason for the difference in E if any viscous effect is present. Stress-strain nonlinearity (plastification, effects of discontinuities, …) could be a reason for discrepancy as well.
Finally, it should be pointed out that assigning statistical properties to the prior information (i.e., mean and variance) is usually a difficult task. Particularly, assigning the weight of the prior information with respect to the measurements during construction is always complex. In some cases it is possible to incorporate those weights as additional parameters to be identified [Led88], [Led96b].

5. Additional related topics

5.1. Capabilities of the Maximum Likelihood Approach

In the previous sections, the Maximum Likelihood Approach and the Gauss-Newton Method have been the fundamental tools to solve the problem of identifying parameters of a defined model from field measurements. The maximum likelihood formulation is a general framework that allows different options when deciding how to pose the inverse problem: it may include the error structure of the measurements, it may include different types of measurements in a consistent manner and also it may include prior information if necessary. Obviously it is always possible to adopt a more traditional approach considering the covariance matrices just as weighing matrices without using the probabilistic framework.

In order to show the capabilities of the formulation, most of the examples included only two parameters and an elastic model. That is, a set of simple examples were preferred to present the basis of the procedure. More sophisticated examples can be found in the literature since late nineties and only some particular citations follow. [Gen96] shows the application of the standard method to the identification of 4 parameters (3 Young's modulus corresponding to 3 materials and *Ko*). A discussion on the reliability of the solution is also presented. [Led97] applies the formulation to the identification of two parameters of the Cam-clay model from displacements and pore water pressures measured in a centrifuge test concerning a tunnel excavation problem. Finally, [deS12] presents the identification of *Ko* and the stiffness of the joints from the lining of a segmental tunnel from London Underground, using the

measurements of segment rotations at different times. In this case a coupled H-M elastic-perfect plastic model (Mohr-Coulomb type) was used to simulate soil behaviour, but the optimization was obtained through simple inspection of the objective function.

The use of the Gauss-Newton algorithm (or any algorithm based on the computation of the derivatives of the objective function) seems to be very robust, but it may become very expensive if the number of parameters involved is large. In this case, methods that only compute the objective function (and not its derivatives) may be more appropriate. Some comments on this topic are included in section 5.4.

5.2. Identification of models

One of the first questions that arise from the above described procedures is whether the model can be identified or not from the field data as well as the parameters. The discrepancy between measurements and computations, assigned to error measurements, is quite often a consequence of an inappropriate model.

In some particular problems it is quite easy to improve a model by incrementing additional "extension" terms [Haf98], but this is more difficult in the context of Geomechanics. Different models in Geotechnical problems may imply moving from elastic to elastoplastic behaviour or changing the geometry itself. That change seems to be too fundamental to be represented by an "extension" term. However, there are promising new "soft computing" techniques in which the model evolves and learns according to the measurements available (i.e. neural network based procedures), [Has10].

A classical and still possible approach consists on comparing objective functions directly: the objective function that gives the minimum value is the one with "less error" between measured and computed variables. If the measurement points are the same and only the model changes, this process allows discriminating between competing models. However, in many practical situations, there is external information about the problem that helps in deciding which model is suitable for each case. Despite that, the comparison between objective functions is always an objective manner of discriminating between models and it can be useful, particularly when interpreting laboratory and field tests or when identifying the geometry.

5.3. Optimal design of experiments

When thinking in terms of "inverse problems" there is always an issue regarding where to install the instrumentation in order to obtain as much information as possible from measurements. The classical idea is to measure where we expect the maximum movements or the maximum pore water pressure changes, but quite often it is difficult to know that in advance.

The procedure described here provides information on the quality of the parameters identified. From expression (15) it becomes clear that we should minimize the variances of the identified parameters in order to obtain the maximum information from the measurements. That requires large values of the sensitivity matrix, **A**, which means that measurements are highly dependent on the parameters. Therefore, an optimum design of a field test or any experiment would require evaluating **A**, looking for measurement points where the sensitivity is maximized.

There are different criteria to define that maximum, i.e., we could maximize the determinant of matrix $\mathbf{A}^t\mathbf{A}$ [Haf98] or a combination of its terms that has statistical meaning [Fin05]. The key aspect, however, is that sensitivity matrix depends in general on the parameters, so the maximum sensitivities are obtained at different points according to the values of the parameters. As an example, the optimal points for measuring displacements may be different depending on soil stiffness [Mur88], [Xia03].

In some particular cases, it is possible to evaluate the sensitivity matrix and to check its tendency or just check if any term may become always zero. It has been shown [Led03] that in the context of tunneling and for elastic models, measurements in any point at 45 degrees provides the larger error (assuming vertical and horizontal directions as principal stress directions). Then, it is possible to decide where not to measure in this particular geometry.

It can be concluded that the design of any experiment can be, to some extent, optimized to obtain the maximum information, but it is necessary to know in advance the range of the values of the parameters.

5.4. Alternative approaches

The number of contributions dealing with inverse problems in Geomechanics has increased significantly since 1980, when the initial works were published [Gio80]. Before that, back-analysis was performed in an "ad hoc" manner. Many of the papers used the Least Squares criterion and the Gauss-Newton algorithm for optimization purposes, although other approaches were also developed: Kalman Filter approach [Mur88], Bayesian approach [Civ83], and the described maximum likelihood approach. During the last two decades, some published works improved the optimization procedure in several aspects as in [Cal04], where the parameters that can be identified simultaneously are defined according to its correlation (i.e., parameters highly correlated should not be identified simultaneously). This particular aspect improves the uniqueness of the solution.

In recent years, there is an increasing amount of works using "soft computing" techniques as optimization procedures. This is a general label used for heuristic based techniques including genetic algorithms, [Lev09], [Lev10], neural networks, [Obr09], [Has10], and particle swarm optimization, [Zha09], among others. General-

ly, this type of methods does not require computing the derivatives of the objective function, a cumbersome task when the number of parameters is large. To illustrate this type of procedures, an example combining genetic algorithms and gradient based methods is briefly presented in next section.

6. An example of hybrid method

In this section an optimization procedure based on the combination of genetic algorithms and Gauss-Newton method is briefly presented. The procedure is described in more detail in [deS14]. Let us consider again a synthetic case involving a tunnel excavation problem in a homogeneous material as shown in Figure 9. An extensometer measuring vertical displacements, 2 m away from the tunnel side, is also depicted in the Figure.

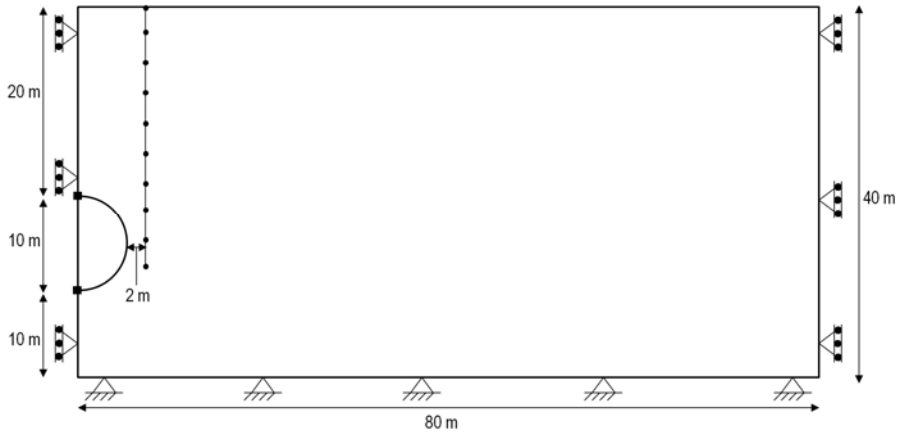


Figure 9. Geometry and measurement points considered for the hybrid analyses.

The code Plaxis [Bri08] was used for the analyses and the soil was assumed to behave according to the Hardening-Soil model [Sch99], one of the constitutive laws implemented in Plaxis. The parameters to identify are the coefficient of lateral earth pressure (K_o) and the reference Young's modulus for unloading and reloading conditions (E_{ur}^{ref}), to the reference pressure (p^{ref}).

A genetic algorithm is an optimization method proposed by [Hol75] and based on Darwin's theory of evolution. An objective function based on a least squares criterion as defined in (5) was adopted. The value of the objective function defines how good an individual is (that is, how good a set of parameters is).

First of all, the range of the parameters (the search space) should be defined and an initial random population is generated. The objective function for each individual is computed and we can check what the best group of individuals is. Then we need a criterion to generate a new population of individuals with the aim of improving their fitness (decreasing their objective function values). That criterion is based on operations called "crossover" and "mutation". After a few iterations creating new sets of populations, it is possible to define a zone in the parameter's space where the minimum is probably located. The method does not guarantee that a global minimum is found, but it is expected to obtain relatively close solutions. Therefore it is convenient to combine the method with a classical gradient method that looks for the minimum in a strict sense.

Table 3 shows the parameters used to generate the measurements. The simulation was carried out in three stages: Phase 1: Tunnel excavation using the Plaxis method ΣMStage to simulate a volume loss close to 0.5% (ΣMStage = 0.2); Phase 2: Tunnel construction activating the lining; Phase 3: Dissipation of all the excess of water pressure caused by the tunnel construction process (consolidation).

Ten points located on the extensometer 2 m away from the tunnel side were considered and the final vertical displacements after all phases corresponding to the values: E_{ur}^{ref}=75000kN/m^2 and K_0=1.5 were considered as "measured" displacements. Figure 10 presents the initial population on the objective function. As only two parameters are involved, it is possible to see directly the evolution of the procedure. After 3 generations and evaluating 120 individuals (that is, 120 direct analyses), it was decided to apply the gradient method. An elliptic zone was defined including all the individuals with an error less than a threshold value. Inside that zone, the gradient method was applied starting from the center of the ellipse. In this particular example the starting point was: E_{ur}^{ref} = 84150kN/m^2 , K_0 = 1.511 (Figure 11).

Table 3. Parameters used to generate the measurements of the tunnel excavation. γ_{unsat}: unsaturated soil weight, γ_{sat}: saturated soil weight, E_{50}^{ref}: secant stiffness in standard drained triaxial test, E_{oed}^{ref}: tangent stiffness for primary oedometer loading, E_{ur}^{ref}: unloading / reloading stiffness, c_{ref}: cohesion, φ: internal friction angle, R_{inter}: interface strength factor and K_0: coefficient of lateral earth pressure. After [deS14].

Parameter	Value
γ_{unsat}	19 [kN/m^3]
γ_{sat}	21 [kN/m^3]
Permeability	0.026 [m/day]
E_{50}^{ref}	25000 [kN/m^2]
E_{oed}^{ref}	20000 [kN/m^2]
E_{ur}^{ref}	50000 - 350000 [kN/m^2]
c_{ref}	10 [kN/m^2]
φ	28 [deg]
R_{inter}	0.6
K_0	0.4 - 2

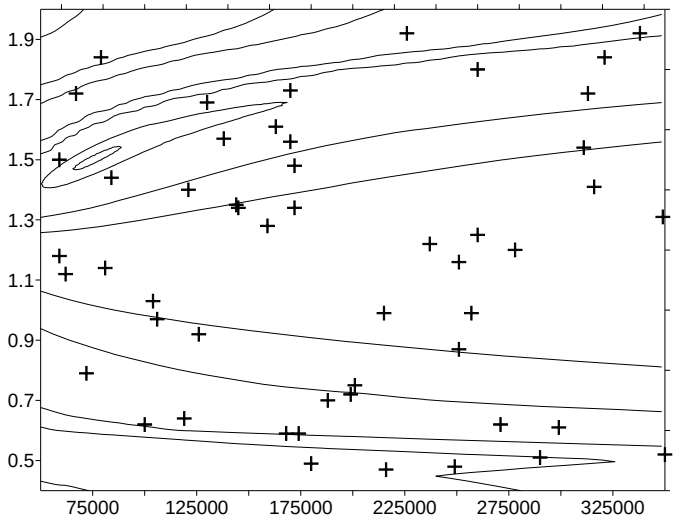


Figure 10. Analysis using genetic algorithm: Initial population. K_0 is represented in the vertical axis and E_{ur}^{ref} [kN/m^2] in the horizontal one. After [deS14].

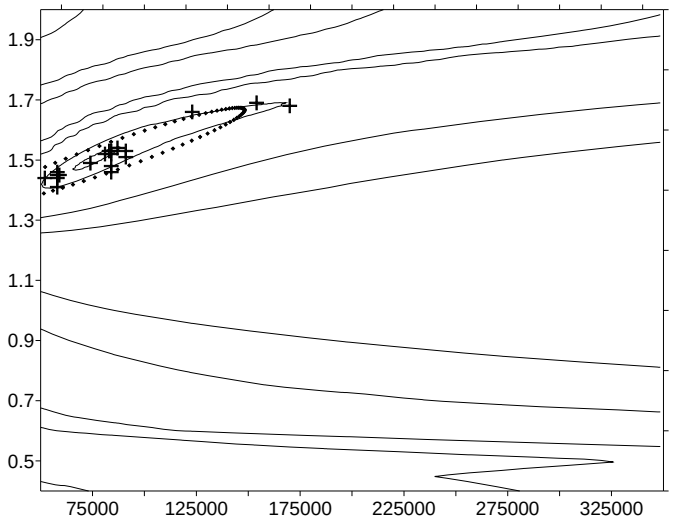


Figure 11. Best individuals from all genetic algorithm generations and new search space (ellipse) for the Gauss-Newton method. After [deS14].

After five additional iterations the Gauss-Newton method reached the correct parameters. In this case the sensitivity matrix was evaluated using a finite difference scheme.

In general it is less expensive, in computational terms, combining both Genetic Algorithm and Gauss-Newton method than using just Genetic Algorithms only. Also this "hybrid" method is more robust than using an individual algorithm only. When the number of parameters increases, this procedure seems to be a promising method to perform back-analysis in a systematic manner.

7. Concluding remarks

A description of the procedures available to carry out an inverse analysis in the context of Geotechnical Engineering has been presented. The Maximum Likelihood Approach has been used as a global framework to describe some fundamental concepts related to these techniques: objective function, minimization algorithms, uniqueness of the solution, sensitivity matrix, the role of error measurements, incorporating prior information and reliability of the parameters identified. Several synthetic examples and a real case have been used to illustrate the capabilities of the formulation and the difficulties encountered when performing back-analysis.

The text also includes some comments on related topics, as the optimal design of experiments or the identification of models in Geomechanics. Also, a list of alternative procedures is presented, leading to the description of a hybrid method, combining genetic algorithms and Gaus-Newton algorithm.

Nowadays the procedures available to perform back-analysis in a systematic manner in Geomechanics are quite mature, but still the methods are not fully used by the Geotechnical community. Most probably, in the near future, the commercial geotechnical finite element codes will incorporate an identification module. It is not the matter of substituting the judgment of an experimented engineer, but the fact that measurements provide valuable information that can be incorporated to our design in a consistent and rational manner.

Acknowledgements

This work has been developed at UPC-BarcelonaTech during the last two decades, and the interaction with my colleagues and PhD students has been fundamental to produce an original research. Particularly, I want to acknowledge Antonio Gens, Eduardo E. Alonso, Enrique Romero and Cristian de Santos for that fruitful interaction.

References

[Bri08] R.B.J. Brinkgreve, W. Broere. *Plaxis 2D Manual (version 9),* Balkema, 2008.

[Bur75] K.V. Bury. *Statistical Models in Applied Science.* Wiley, New York, 1975.

[Cal04] Michele Calvello, Richard J. Finno. Selecting parameters in model calibration by inverse analysis. *Computers and Geotechnics*, 31:411-425, 2004.

[Civ83] A. Cividini, G. Maier, A. Nappi. Parameter estimation of a static geotechnical model using a Bayes' approach. *Int. J. Rock Mech. Mining Sci. Geomech. Abstr.,* 20:215-226, 1983.

[Dav80] E.H. Davis, M.J. Gunn, R.J. Mair, H.N. Seneviratne. The stability of shallow tunnels and underground openings in cohesive material. *Géotechnique,* 30(4):397-416, 1980.

[deS12] C. De Santos, A. Ledesma, A. Gens. Backanalysis of measured movements in ageing tunnels. *Proc. 7th Int. Symosium on Geotechnical aspects of underground construction in soft ground – TC28, 17-19 May 2011*. G. Viggiani ed. CRC press, 223-229, Rome, 2012.

[deS14] C. de Santos, A. Ledesma, A. Gens. Hybrid minimization algorithm applied to tunnel backanalysis. *Proc. 14th Int. Conf.of the Int. Assoc. for Computer Methods and Advances in Geomechanics. Kyoto*, in press. 2014.

[Dun70] J.M. Duncan, C.Y. Chang. Nonlinear analysis of stress and strain in soils. *J. Soil Mech. Found. Engng. ASCE,* 96:1629-1653, 1970.

[Edw72] A.W.F. Edwards. *Likelihood.* Cambridge University Press, Cambridge, UK, 1972.

[Fin05] Richard J. Finno, Michele Calvello. Supported excavations: Observational Method and Inverse Modeling. *J. Of Geotechnical and Geoenvironmental Engineering*, 131(7):826-836, 2005.

[Fle81] R. Fletcher. *Practical Methods of Optimization. 1. Unconstrained optimization. 2. Constrained optimization*. Wiley, Chichester, 1981.

[Gen88] A. Gens, A. Ledesma, E.E. Alonso. Back analysis using prior information. Application to the staged excavation of a cavern in rock. *6th*

Int. Conf. on Num. Methods in Geomechanics. Innsbruck, Balkema, pages 2009-2016, 1988.

[Gen96] A. Gens, A. Ledesma, E.E. Alonso. Estimation of parameters in geotechnical backanalysis – II Application to a tunnel excavation problem. *Computers and Geotechnics, 18(1):29-46*, 1996.

[Gio80] G. Gioda, G. Maier. Direct search solution of an inverse problem in elastoplasticity: identification of cohesion, friction angle and in situ stress by pressure tunnel tests. *Int. J. Numer. Meth. Engng., 15:1823-1848, 1980*

[Haf98] Raphael T. Haftka, Elaine P. Scott, Juan R. Cruz. Optimization and experiments: A survey. *Applied Mechanics Reviews, ASME*, 51(7):435-448, 1998.

[Has10] Youssef M.A. Hashash, Séverine Levasseur, Abdolreza Osouli, Richard Finno, Yann Malecot. Comparison of two inverse analysis techniques for learning deep excavation response. *Computers and Geotechnics*, 37:323-333, 2010.

[Hol75] J.H. Holland. *Adaptation in Natural and Artificial Systems: An Intoductory Analysis with Applications to Biology, Control, and Artificial Intelligence.* Univ. of Michigan Press, Ann Arbor, 1975.

[Led88] A. Ledesma, A. Gens, E.E. Alonso. Estimation of parameters from instrumentation data obtained during the staged excavation of a cavern. *Rock Mechanics and Power Plants. M. Romana, ed*. Balkema, Rotterdam, vol. 2, 1988

[Led91] A. Ledesma, A. Gens, E.E. Alonso. Identification of parameters of nonlinear geotechnical models. *7th Int. Conf. Computer Methods and Advances in Geomechanics. Beer, Booker & Carter eds.,* Balkema, Rotterdam, pages 1005-1010, 1991.

[Led96a] A. Ledesma, A. Gens, E.E. Alonso. Estimation of parameters in geotechnical backanalysis – I Maximum Likelihood Approach. *Computers and Geotechnics, 18(1):1-27*, 1996.

[Led96b] A. Ledesma, A. Gens, E.E. Alonso. Parameter and variance estimation in Geotechnical Backanalysis using prior information. *Int. J. for Numerical and Analytical Methods in Geomechanics, 20:119-141, 1996.*

[Led97] A. Ledesma, A. Gens. Inverse analysis of a tunnel excavation problem from displacement and pore water pressure measurements. *Material Identification using mixed numerical and experimental methods. Proceedings of the Euromech 357 coloquium, H. Sol & C.W. Oomens eds.*, Kluwer, Dordrecht, 163-172, 1997.

[Led03] A. Ledesma. Análisis retrospectivos sistemáticos durante la excavación de túneles. *Jornadas Hispano-Lusas sobre Obras Subterráneas: Relevancia de la prospección y observación geotécnicas,* Madrid, 15-16 Sept., CEDEX, 441-448, 2003

[Lev09] S. Levasseur, Y. Malécot, M. Boulon, E. Flavigny. Statistical inverse analysis based on genetic algorithms and principal component analysis: Method and developments using synthetic data. *Int. J. for Numerical and Analytical Methods in Geomechanics, 33:1485-1511,* 2009.

[Lev10] S. Levasseur, Y. Malécot, M. Boulon, E. Flavigny. Statistical inverse analysis based on genetic algorithms and principal component analysis: Applications to excavation problems and pressurometer tests. *Int. J. for Numerical and Analytical Methods in Geomechanics, 34:471-491,* 2010.

[Mar63] D.W. Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *J. Soc. Indust. Appl. Math., 11:431-441, 1963.*

[Mur88] Akira Murakami, Takashi Hasegawa. Back analysis using Kalman filter-finite elements and optimal location of observed points. *6th Int. Conf. on Num. Methods in Geomechanics. Innsbruck,* Balkema, pages 2051-2058, 1988.

[Obr09] Rafal F. Obrzud, Laurent Vulliet, Andrzej Truty. A combined neural network/gradient-based approach for the identification of constitutive model parameters using self-boring pressurometer tests. *Int. J. for Numerical and Analytical Methods in Geomechanics*, 33(6):817-849, 2009.

[Puz10] A.M. Puzrin, E.E. Alonso, N.M. Pinyol. *Geomechanics of failures.* Springer, Dordrecht, 2010.

[Sch99] T. Schanz, P.A. Vermeer, P.G. Bonnier. Formulation and verification of the hardening soil model. *Beyond 2000 in Computational Geotechnics*. R.B.J. Binkgreve ed., Balkema, Rotterdam, 281-290, 1999.

[Tar87] A. Tarantola. *Inverse Problem Theory.* Elsevier, Amsterdam, 1987

[Ter67] Karl Terzaghi, Ralph B. Peck. *Soil Mechanics in Engineering Practice.* 2nd edition, John Wiley & Sons, New York, 1967.

[Wal90] E. Walter, L. Pronzato. Qualitative and quantitative experiment design for phenomenological models – A survey. *Automatica*, 26:195-213, 1990.

[Wig72] R.A. Wiggins. The general linear inverse problem: implication of surface waves and free oscillations for earth structure. *Reviews of Geophysics and Space Physics*, 10(1):251-285, 1972.

[Xia03] Z. Xiang, G. Swoboda, Z. Cen. *Int. Journal of Geomechanics, ASCE,* 3(2):205-216, 2003.

[Zha09] H. Zhao, S. Yin. Geomechanical parameters identification by particle swarm optimization and support vector machine. *Applied Mathematical Modelling*, 33(10):3997-4012, 2009.

Calibration of soil constitutive laws by inverse analysis

Michele Calvello

Università di Salerno, Italy

When a model is calibrated by iteratively changing the estimates of the model input parameters until the value of an objective function, which quantifies the match between observed and computed results, is minimized we are dealing with inverse analysis. The major advantage of an inverse modelling is the automatic and objective calculation of the parameter values that produce the best fit between measured data (often called observations) and computed results. The main difficulties are related to the complexity of most numerical models, which sometimes cause problems of non-uniqueness and instability of the solution or insensitivity of the results to changes in the values of the parameters. This chapter presents the main aspects related to inverse analysis techniques used to calibrate the parameters of soil constitutive laws. It comprises three main sections respectively dealing with: a computer code designed to allow inverse modeling posed as a parameter estimation problem; the use of inverse analysis to calibrate soil models from results of laboratory experiments; an inverse analysis procedure to update the design predictions of a supported excavation system using monitoring data collected during construction.

1 Introduction

Model calibration means "tuning" the parameters and/or the components of a given model so that values measured within a real system (e.g., results of laboratory or field tests, monitoring data from engineering projects) are matched by equivalent computed values until the resulting calibrated model accurately represents the main aspects of the system. Despite their apparent utility, inverse analysis techniques are used, to this purpose, much less than expected and most typically numerical models are calibrated using trial-and-error methods. With an inverse modeling approach, a model is calibrated by iteratively changing the estimates of the model input parameters until the value of an objective function, which quantifies the match between observed and computed results, is minimized. Inverse analysis works in the same way as a non-automated calibration approach: parameter values and other aspects of

the model are adjusted until the model's computed results match the observed behavior of the system. However, use of an inverse model provides additional results and statistics that offer numerous advantages in model analysis and, in many instances, expedites the process of adjusting parameter values [Cal02]. The fundamental benefit of inverse modeling is its ability to automatically calculate parameter values that produce the best fit between observed and computed results. In addition other benefits are derived, including: substantial time saving over traditional trial-and-error calibration methods; statistics that quantify quality of calibration, data shortcomings and reliability of parameter estimates and predictions; identifications of issues that are easily overlooked during non-automated calibration. The main difficulties inherent to inverse modeling algorithms are related to the complexity of real systems, almost always modeled non-linearly, which sometimes leads to problems of: insensitivity, when the observations do not contain enough information to support estimation of the parameters; non-uniqueness, when different combination of parameter values match the observations equally well; instability, when slight changes in model variables radically change inverse model results.

2 A model independent inverse analysis algorithm: UCODE

UCODE [Poe98] is a computer code designed to allow inverse modeling posed as a parameter estimation problem. UCODE can be effectively used in geotechnical modeling because it works with any application software that can be executed in batch mode. Its model-independency allows the chosen numerical code to be used as a "closed box" in which modifications only involve model input values. Figure 1 shows a detailed flowchart of the parameter optimization algorithm used in UCODE. Note that the minimization requires multiple runs of the finite element code.

The weighted least-squares objective function S(b) is expressed by:

$$S(\underline{b})=\left[\underline{y}-\underline{y}'(\underline{b})\right]^T \underline{\omega}\left[\underline{y}-\underline{y}'(\underline{b})\right]=\underline{e}^T\underline{\omega}\,\underline{e} \tag{1}$$

where: b is a vector containing values of the number of parameters to be estimated; y is the vector of the observations being matched by the regression; y'(b) is the vector of the computed values which correspond to observations; ω is the weight matrix; e is the vector of residuals.

Non-linear regression is an iterative process. The modified Gauss-Newton method used to update the input parameters is expressed as:

$$\left(\underline{C}^T\underline{X}_r{}^T\,\underline{\omega}\,\underline{X}_r\,\underline{C}+\underline{I}m_r\right)\underline{C}^{-1}\underline{d}_r=\underline{C}^T\underline{X}_r{}^T\,\underline{\omega}\left(\underline{y}-\underline{y}'(\underline{b}_r)\right) \tag{2}$$

$$\underline{b}_{r+1}=\rho_r\underline{d}_r+\underline{b}_r \tag{3}$$

where: $\underline{d}_r$ is the vector used to update the parameter estimates $\underline{b}$; r is the parameter estimation iteration number; $\underline{X}_r$ is the sensitivity matrix ($X_{ij}=\partial y_i/\partial b_j$) evaluated at parameter estimate $\underline{b}_r$; $\underline{C}$ is a diagonal scaling matrix with elements c_{jj} equal to $1/\sqrt{(\underline{X}^T \underline{\omega}\ \underline{X})_{jj}}$; $\underline{I}$ is the identity matrix; m_r is a parameter used to improve regression performance; ρ_r is a damping parameter.

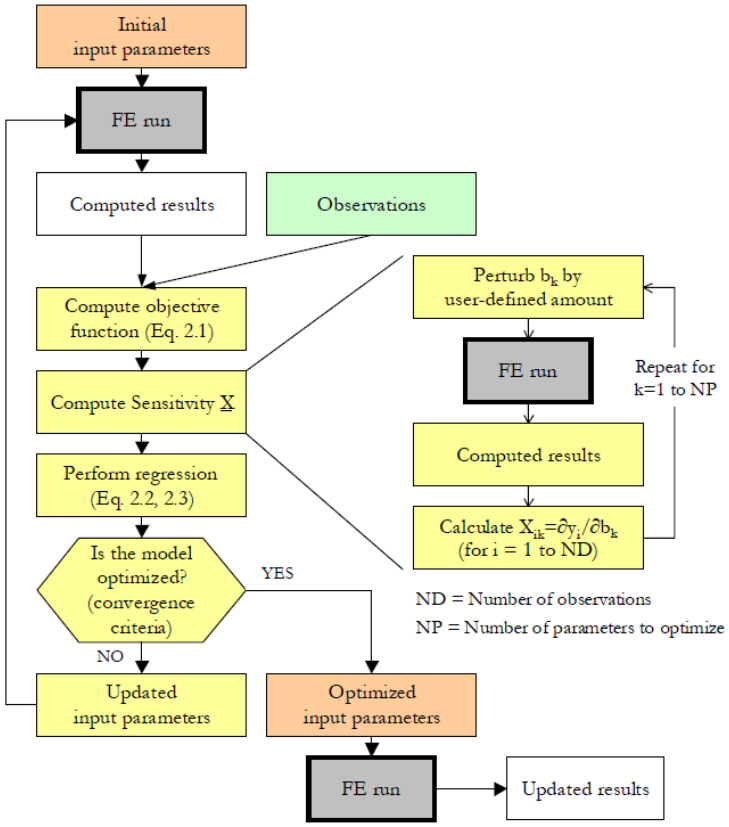


Figure 1: Parameter optimization algorithm flowchart [Fin05].

Multiple runs of the FE model are required to update the input parameters at a given iteration because the sensitivity matrix $\underline{X}_r$ is computed using a perturbation method. At any iteration every input parameter $\underline{b}_k$ is independently perturbed by a fractional amount to compute the results' response to its change. Sensitivities are calculated by forward or central differences approximations. For these approximations each iteration requires (NP+1) and (2NP+1) runs, respectively, to estimate a new set of updated parameters, where NP is the number of parameters optimized simultaneously. Computation time may become an issue for very complicated finite element models, depending on how much time is needed for a single model run. At a given iteration, after performing the modified Gauss-Newton optimization (Eq. 2 and 3), UCODE

decides whether the updated model is optimized according to two convergence criteria. The parameter estimation is said to converge if either: i) the maximum parameter change of a given iteration is less than a user-defined percentage of the value of the parameter at the previous iteration; ii) the objective function, S($\underline{b}$), changes less than a user-defined amount for three consecutive iterations. When the model is optimized the final set of input parameters is used to run the model one last time and produce final "updated" results.

2.1 Model fit statistics

Different quantities can be used to evaluate the model fit. At first, one can consider the magnitude of the weighted and unweighted residuals and their distribution both statistically and relative to independent variable values such as location and time. In initial model runs large residuals and weighted residuals can indicate gross errors in the model, in the data, in how the observed quantity is simulated, or in the weighting. In subsequent model runs, after the gross errors have been corrected, other statistics become increasingly important. A commonly used indicator of the overall magnitude of the weighted residuals is the model error variance, s^2:

$$s^2 = \frac{S(\underline{b})}{ND-NP} \tag{4}$$

where: S($\underline{b}$) is the objective function; ND is the number of observations; NP is the number of estimated parameters.

The value of the objective function (Eq.1) is also used to indicate model fit informally, because its variation indicates by how much an optimized model improves with respect to the initial simulation of a problem. Graphical analyses are also useful to assess the validity of the model optimization. Ideally, weighted residuals are scattered evenly about 0.0, and their size is not related to the simulated values. Weighted residuals plotted on maps or time graphs should not show any discernible patterns and should appear random. When weighted observations are plotted against weighted simulated values ideally the points should fall close to a line, with slope equal to 1.0 and an intercept of zero, and the correlation coefficient between the two series should be close to 1.0.

2.2 Parameter statistics

The relative importance of the input parameters being simultaneously estimated can be defined using parameter statistics, such as: the sensitivity of the predictions to changes in parameters values, the variance-covariance matrix, confidence intervals and coefficients of variation. To evaluate the sensitivity of the predictions to parameters changes, it is useful to investigate one percent sensitivities, dss_{ij}, scaled sensitivities, ss_{ij}, and composite scaled sensitivities, css_j:

$$dss_{ij} = \frac{\partial y_{i'}}{\partial b_j}\frac{b_j}{100} \tag{5}$$

$$ss_{ij} = \left(\frac{\partial y_{i'}}{\partial b_j}\right) b_j {\omega_{ii}}^{1/2} \tag{6}$$

$$css_j = \left[\sum_{j=1}^{ND}\left(\left(\frac{\partial y_{i'}}{\partial b_j}\right) b_j {\omega_{ii}}^{1/2}\right)^2\Bigg|_{\underline{b}} /ND\right]^{1/2} \tag{7}$$

where: y'_i is the i^{th} simulated value; y_i/b_j is the sensitivity of the i^{th} simulated value with respect to the j^{th} parameter; b_j is the j^{th} estimated parameter; ω_{jj} is the weight of the i^{th} observation.

One percent scaled sensitivities represent the amount that the simulated value would change if the parameter value increased by one percent. Scaled sensitivities are dimensionless quantities that can be used to compare the importance of different observations to the estimation of a single parameter or the importance of different parameters to the calculation of a simulated value. Composite scaled sensitivities indicate the total amount of information provided by the observations for the estimation of one parameter.

The reliability and correlation of parameter estimates can be analyzed by using the variance-covariance matrix, $\underline{V}(\underline{b}')$, for the final estimated parameters, $\underline{b}'$, calculated as:

$$\underline{V}(\underline{b}') = s^2\left(\underline{X}^T\ \underline{\omega}\ \underline{X}\right)^{-1} \tag{8}$$

where: s^2 is the error variance; $\underline{X}$ is the sensitivity matrix; $\underline{\omega}$ is the weight matrix.

The diagonal elements of matrix $V(\underline{b}')$ equal the parameter variances, the off-diagonal elements equal the parameter covariances. Parameter variances and covariances are most useful when used to calculate other statistics: confidence intervals for parameter values, CI; coefficients of variation, cov_i; correlation coefficients cor(i,j):

$$CI: b_j \pm \mathrm{t}\left(n, 1.0 - {\alpha}/{2}\right)\sigma_{b_j} \tag{9}$$

$$cov_i = {\sigma_i}/{b_i} \tag{10}$$

$$cor(i,j) = cov(i,j)/\left(\sqrt{var(i)}\sqrt{var(j)}\right) \tag{11}$$

where: $t(n,1.0-\alpha/2)$ is the student-t statistic for n degrees of freedom and a significant level of α; σ_{bj} is the standard deviation of the j^{th} parameter. σ_i is the standard deviation of parameter b_i, cor(i,j) indicate the correlation between the i^{th} and j^{th} parameter; cov(i,j) equal the off-diagonal elements of $\underline{V}(\underline{b}')$; var(i) and var(j) refer to the diagonal elements of $\underline{V}(\underline{b}')$.

A confidence interval is a range that has a stated probability of containing the true value of the estimated variable. The width of the confidence interval can be thought of as a measure of the likely precision of the estimate, with narrow intervals indicating greater precision. The coefficients of variation provide dimensionless numbers with which the relative accuracy of different parameter estimates can be compared. Correlation coefficients close to –1.0 and 1.0 are indicative of parameters that cannot be uniquely estimated with the observations used in the regression.

2.3 Observations' weighting

The weights assigned to the observations are an important part of the regression analysis because they influence the value of the objective function, and thus the regression results. UCODE uses a diagonal weight matrix. Weighting is used to reduce the influence of observations that are less accurate and increase the influence of observations that are more accurate. For problems with more than one kind of observation, weighting also produces weighted residuals that have the same units, so that they can be squared and summed. The weight of every observation, ω_{ii}, is equal to the inverse of its error variance, σ_i^2:

$$\omega_{ii} = 1/{\sigma_i}^2 \tag{12}$$

Users assign the weight of an observation by specifying a value for its variance, standard deviation or coefficients of variation. At the end of the regression analysis, the value of the model error variance, s^2 (Eq.4), can be used to evaluate the consistency between the model fit and the measurement errors, as expressed by the observations' weights. Values larger than 1.0 indicate that the model fits the data less well than would be accounted for by expected measurement error.

2.4 On constraining parameters during optimization

While limiting constraints on parameter values may, at times, appear to be necessary, UCODE users are not allowed to set upper or lower limits on parameters to be estimated, as this practice might disguise model inaccuracies [Hil98]. Indeed, unrealistic optimized input parameter values are likely to indicate either a more fundamental model error (thus, users are prompted to find and correct the error) or observations not containing enough information to estimate the parameters. Responses to the second circumstance could be: the exclusion of the parameter from the optimization, the use of prior information on the parameter value. Using prior information allows direct measurements of model input parameters to be included in the regression and tends to produce estimates that are close to specified parameter values. The effect that the prior information has, in "forcing" an estimated parameter to remain close to a specified value, depends from its weight. Users must treat the prior information like an extra observation point. Its influence in conditioning the response of

the regression analysis depends on both the number of observations included and the weight of the prior information relative to the weight of the other observations. A problem with the "unconstrained approach" used by UCODE is that many geotechnical parameters have natural constrains. Many soil models' input parameters, for instance, admit only positive values (e.g. Young's modulus, cohesion), and some of them are bounded by upper and lower physical limits (e.g. Poisson's ratio, friction angle). To address the first problem UCODE allows the user to optimize the log-transformed value of the native parameter. This produces an inverse problem that prevents the actual parameter value from becoming negative. Nothing is directly implemented in UCODE to address the second problem. However, users can specify functions of the parameter values to be used as input to the application model. Users can thus relate the input parameter to a bounded "mapping function." For example, the input parameter to be constrained, x, can be mapped by a hyperbolic function expressed in terms of e^x. Figure 2 shows such a function, its expression given by:

$$f(x) = y_1 + e^x / \left({}^{e^x}/_{(y_2 - y_1)} + {}^{1}/_{Tan} \right) \tag{13}$$

where: x is the native parameter, y_1 is the lower limit of x, y_2 is the upper limit of x, and Tan is the initial tangent in the y-e^x space.

To "center" the mapping function around a specific value y_0, Tan must be equal to:

$$Tan = (y_0 - y_1)(y_2 - y_1)/(y_2 - y_1) \tag{14}$$

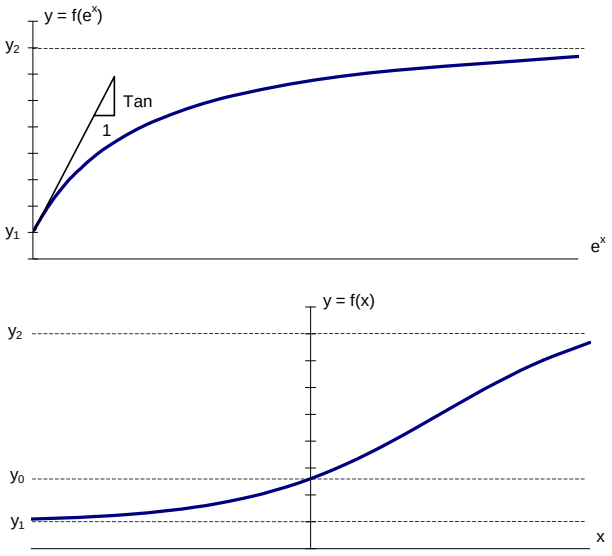


Figure 2: Hyperbolic mapping function to constrain an input parameter value.

3 Calibration of modified cam-clay model from triaxial test results

This section deals with inverse analysis used to calibrate modified cam-clay (MCC) parameters from results of one drained and one undrained triaxial compression tests on specimens of compressible Chicago glacial clays [Cal02]. This approach couples a finite element code and the inverse analysis algorithm UCODE to minimize the differences between computations of stress-strain response and experimental data. The soil specimens used for the laboratory tests are undisturbed samples of lightly overconsolidated low-to-medium plastic clays from downtown Chicago [Fin02]. The MCC model [Ros58] is an isotropic, work hardening, non-linear, elasto-plastic model. The responses are defined in terms of three state variables: the mean normal stress p′, the deviatoric stress q, and the void ratio e. The first two variables are defined as:

$$p' = I_1/3 \tag{15}$$

$$q = \sqrt{3\,J_2} \tag{16}$$

where: $I_1=\sigma_{11}+\sigma_{22}+\sigma_{33}$ is the first Cauchy stress invariant, and $J_2=1/6[(\sigma_{11}-\sigma_{22})^2+(\sigma_{11}-\sigma_{33})^2+(\sigma_{22}-\sigma_{33})^2+6\sigma_{12}+6\sigma_{13}+6\sigma_{23}]$ is the second deviatoric stress invariant.

The four MCC input parameters are: λ, κ, M and G. Table 1 shows their meaning and the conventional way of estimating them. Parameters λ and κ define the model hardening law, M locates the Critical State Line (CSL) in the p′-q space, and G, along with κ, defines the elastic behavior inside the yield surface. Beside the initial values of the state variables, the initial conditions include a parameter expressing the stress history of the soil, p_c*, and the critical state void ratio, e_{cs}, defining the position of the CSL in the e-p′ space. The initial estimates of κ and λ are based on results from consolidation tests; M is estimated assuming a straight failure line passing through zero in p′-q space; G is estimated by averaging the secant shear stiffness at a shear stress level of 50% of the failure value. The initial value of p_c*, assuming the soil OCR ≅ 1, is set equal to the consolidation stress of the test modeled, and e_{cs} is estimated as the value of the CSL at p′=1 kPa.

Table 1: Modified Cam Clay input soil parameters to optimize.

Parameter	Meaning	Initial estimate
λ	Slope of rebound isotropic consolidation curve	C_r / 2.303
κ	Slope of virgin isotropic consolidation curve	C_c / 2.303
M	Slope of the failure line in q-p' space	6 sin φ / (3 – sin φ)
G	Shear modulus	q/γ at 50% $q_{failure}$

Two triaxial compression tests form the basis of the optimization shown herein: one performed in drained conditions (D1) and one in undrained conditions (U1). Both specimens were isotropically consolidated and then sheared by increasing the vertical principal stress to failure. The finite element code JFEST was used to simulate the triaxial tests [Fin91]. The behavior of the samples was considered elemental, thus one single 8-noded isoparametric element was used to model the specimen. To evaluate the match between the soil response and the FE predictions, two curves were used to calibrate the objective function for each type of test: the changes in principal stress difference (q-ε_a) and the volumetric changes (ε_v-ε_a) with axial strain for the drained test D1; the changes in principal stress difference (q-ε_a) and pore pressures (u-ε_a) with axial strains for the undrained test U1. Figure 3 shows the experimental results and the observation points used for the drained and undrained stress-strain curves respectively. The stress-strain curves of the drained test were discretized by considering one observation point every 0.5% axial strain up to ε_a = 12%. Curves for the undrained test were discretized using one observation point every 0.15% axial strain up to ε_a = 4.5%. Therefore, 108 observations were used to calibrate the MCC model. The weight of the observations was assigned through coefficients of variation equal to 5%.

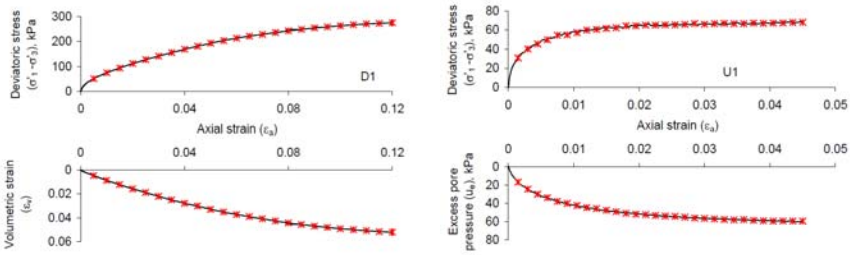


Figure 3: Discretization of experimental results for tests D1 and U1.

Visual examination of the stress-strain plots provides the simplest way to evaluate the fit between experimental and modeled response and thus evaluate the effectiveness of the inverse analysis. Figure 4 shows a comparison between experimental data and computed results when the initial estimates of soil parameters are used to simulate the model response. The computed slightly underpredict the deviatoric stress response of the drained test, and they are not able to capture the initial response of the undrained test. The behavior of the soil, as predicted by the numerical simulation, is softer that the real behavior of the clay sample. Figure 5 shows a comparison between experimental data and computed results when the optimized estimates of soil parameters are used to simulate the model response. The computed results match the overall test results extremely well for both tests at either small or large strain levels. Table 2 shows the values of the four MCC input parameters before and after the optimization. Only small changes in values of κ, λ and M are needed to obtain the best-fit values, whereas the optimized value of G is almost 3 times larger than the initial one. Note that four MCC input parameters are not opti-

mized independently and that the same set of parameters is used to simulate the two tests. Results of the optimization are consistent with what a "knowledgeable" geotechnical engineer would have guessed by looking at Figure 4. The initial set of input parameters underestimates the stiffness of the soil samples, which in the MCC model is mainly (but not uniquely) expressed by parameter G. Indeed G was the parameter that varied the most. However, one could not say a priori by how much the value of G was underestimated in the initial predictions. Moreover, it is very doubtful that one could have estimated by trial-and-error the small fractional change of the first three parameters to arrive at the fit illustrated in Figure 5. Note that if one tries to optimize the value of the stiffness parameter only, keeping the other parameters to their initial value, the fit between experimental data and computed results never become as good [Cal02].

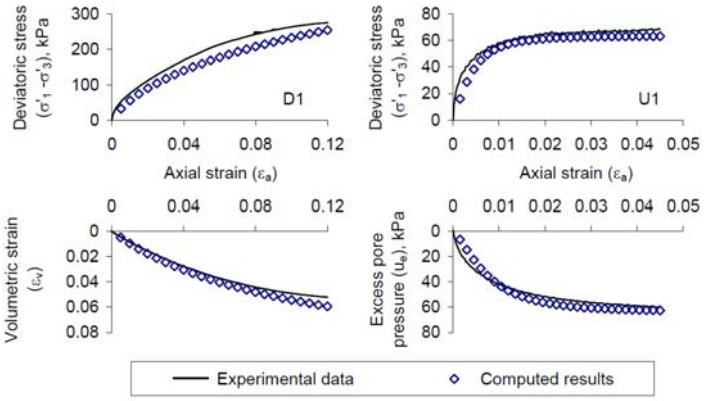


Figure 4: Visual fit between experimental data and computed results for initial estimates of soil parameters.

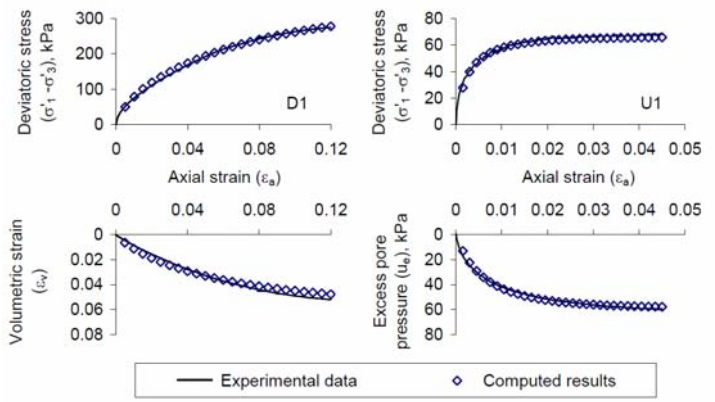


Figure 5: Visual fit between experimental data and computed results for optimized estimates of soil parameters.

Table 2: Input parameter values before and after optimization.

	Input parameter				Visual fit	
	λ	κ	M	G (kPa)	D1 (q-εv)	U1 (q-u)
Initial	0.017	0.130	1.09	4000	marginal good	marginal marginal
After optimization	0.021	0.096	1.04	11054	good good	good good

Table 3 shows a summary of the values of statistics indicating model fit. The statistics clearly show that the optimization of the input parameters improves the fit significantly. Indeed, the objective function value decreased by almost 90%, the model variance is less than 1.0, indicating that the fit is consistent with the error of the observations as expressed by the weighting, and the correlation coefficient is very close to 1.0. Before the optimization the same statistics show that the fit between experimental data and simulated response is inadequate.

Figure 6 shows, in a graphical format, the values of the four MCC input parameters before and after the optimization as well as, on the secondary y-axis, the composite scaled sensitivities of the input parameters, css_j (Eq. 7). For nonlinear problems, the sensitivity is different for different input values, thus css_j are plotted for the initial and the optimized parameters values. Results, however, show that the difference between sensitivity values at the beginning and at the end of the regression analysis is not significant. This indicates that, despite the non-linearity of the model, the influence of a given parameter on the results is relatively constant. If, during the iterative regression analysis, the sensitivities were to vary too much from iteration to iteration it is doubtful that the modified Gauss-Newton optimization method used in UCODE would have been so efficient. Composite scaled sensitivities vary from values smaller than 4 for parameters κ and G, to values larger than 7 for parameter λ and 15 for parameter M. These values indicate that changes of M have the largest impact on the computed values (the higher the sensitivity value of a parameter, the more impact that parameter has on the computed results). This was to be expected since the compression tests were conducted on specimens of lightly overconsolidated clays reconsolidated in the laboratory to stresses greater than or equal to the field value of vertical effective stress. This loading history establishes the stress at the start of the shearing portion of the test very close to or at yield, and hence the parameters associated with plastic hardening and failure (M and λ) would have the most effect on the computed results. However, this does not mean that one can exclude the least sensitive parameters from the regression analysis. Results show that the sensitivities relative to different parameters are all within the same order of magnitude. This indicates that all parameters have a quantifiable effect on the modeled results. Indeed, the least sensitive parameters, G, is the one that changes the most to reach its optimal value.

Figure 7 shows the values of some of the parameter statistics derived from the variance-covariance matrix. A bar chart is used to indicate by what percentage the opti-

mized parameter values changed compared to their initial estimate. The plot also shows the parameters' 95% confidence intervals (Eq. 9). Confidence intervals results indicate that estimates of the κ, λ and M values calculated by the regression algorithm are accurate.

Finally, Table 4 shows the values of the correlation coefficients for the four input parameters (Eq. 11). All the values are very far from either 1.0 or –1.0, indicating that none of the parameters is correlated to any other one. This trend suggests that the observations used in the regression provide enough information for the four parameters to be simultaneously estimated.

Table 3: Input parameter values before and after optimization.

	Model fit statistics			
	Objective function	Model variance	Correlation coefficient	Iterations
Initial	789	7.59	80.1%	
After optimization	93.8	0.90	97.5%	5

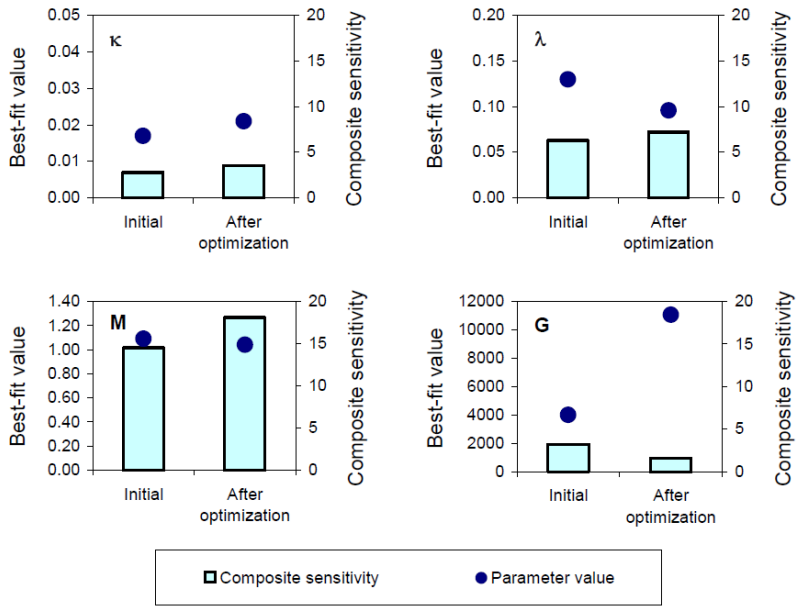


Figure 6: Input parameters value and their composite scaled sensitivity before and after optimization.

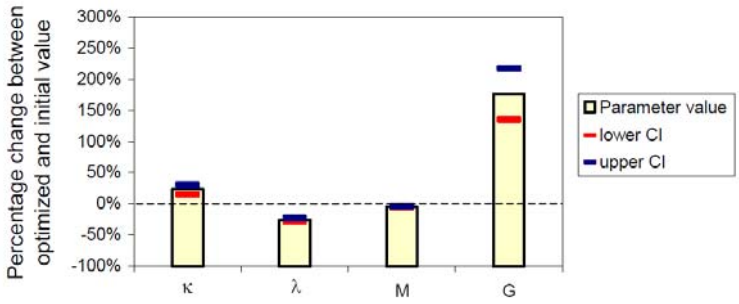


Figure 7: Percentage change between optimized and initial parameter values and confidence intervals.

Table 4: Correlation coefficients among input parameters.

	Correlation coefficients			
	λ	κ	M	G
λ	1	.374	-.035	.451
κ	.374	1	.539	.150
M	-.035	.539	1	-.372
G	.451	.150	-.372	1

In this section the inverse analysis algorithm UCODE was used to calibrate, from triaxial test results, the four input parameters of the MCC soil model. Inverse analysis calibration proved to be more effective in defining these model parameters than conventional estimation methods. Indeed, the use of an optimization algorithm to calibrate the MCC model yielded:

1. almost perfect fit between experimental results and the response computed from the finite element simulation of the tests;
2. objective estimation of the model input parameters;
3. useful model fit statistics, which can be used to evaluate the adequacy of the soil model to simulate the experimental soil response;
4. very valuable parameter statistics, which can help geotechnical engineers in interpreting the features of a soil model.

The most important parameter statistic is probably the composite scaled sensitivity, a powerful statistical measure to detect the parameters that most affect the test results.

4 Update design predictions using monitoring data: a supported excavation case study

4.1 Inverse analysis and the "observational method"

For many geotechnical engineering projects, especially in urban environments, it is critically important to predict the magnitude and distribution of the ground movements caused by the construction procedures. In such cases, a monitoring program is generally planned to record, during construction, ground movements and/or movements of adjacent structures. The monitoring data can be used to evaluate how well the actual construction process is proceeding in relation to the predicted movements as well as to control the construction process and update predictions of movements given the measured deformations at early stages of constructions. The procedure to accomplish the latter task is usually referred to as the "observational method" [Pec69], a framework wherein construction procedures and details are adjusted based upon observations and measurements made as construction proceeds. Employing observed movements in a timely enough fashion to be of practical use in a typical project is generally a difficult task. To obtain inclinometer or optical survey data, process it, and use it to "calibrate" the results of a numerical model of the geotechnical system is a time-consuming process and thus, without employing inverse analysis techniques, this updating process can be done in a timely fashion only with the commitment of significant human and economic resources.

As already discussed in the previous sections, inverse analysis algorithms allow the simultaneous calibration of multiple input parameters. However, identifying the important parameters to include in the inverse analysis can be problematic and, in most practical problems, it is not possible to use the regression analysis to estimate every input parameter of a given model. The number and type of input parameters that one can expect to estimate simultaneously depend upon many factors, including the characteristics of the selected soil model, how the model parameters are combined within the element stiffness matrix in a finite element formulation, the site stratigraphy, the number and type of observations available, the characteristics of the simulated system, and computational time issues. Figure 8 shows a procedural flowchart that can be used for the identification of the soil parameters to optimize with inverse analysis algorithms [Cal04]. The total number of parameters can be reduced, in three steps, to the number of parameters that are likely to be optimized successfully. In Step 1, the number of relevant and uncorrelated parameters of the constitutive model chosen to simulate the soil behavior is determined. The number depends upon the characteristics of the model, the type of observations available, and the stress conditions in the soil. Composite scaled sensitivity values can provide valuable information on the relative importance of the different input parameters of a given model. Parameter correlation coefficients can be used to evaluate which parameters are correlated and are, therefore, not likely to be estimated simultaneously. In Step 2, the number of soil layers to calibrate and the type of soil model used to simulate the layers determine the total number of relevant parameters. An additional

sensitivity analysis may be necessary, within this step, to check for correlations between parameters relative to different layers. In Step 3, a final reduction of the number of parameters to optimize simultaneously may be prompted by the total number of observations available and computational time issues.

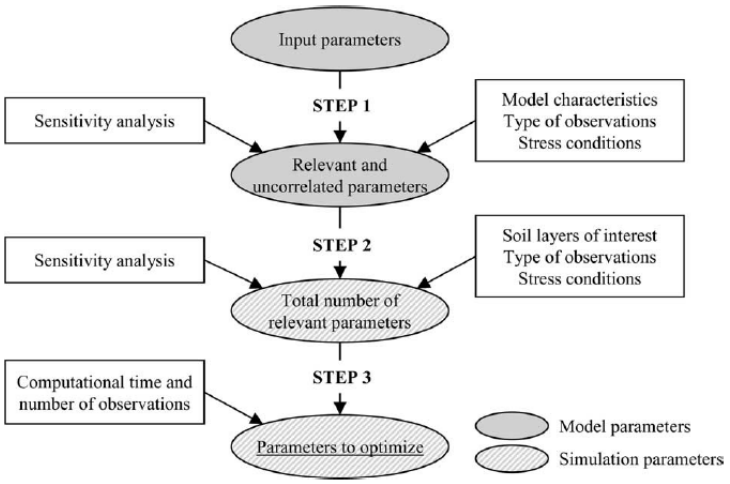


Figure 8: Identification of soil parameters to optimize by inverse analysis [Cal04].

This section shows how inverse analysis based on displacement monitoring data can be used to objectively update the predicted performance of supported excavation systems. To this aim, movements of the soil surrounding an excavation are recorded by inclinometers, which measure lateral deformations at various depths at discrete locations around the construction site. Within the proposed inverse analysis procedure, the recorded data are used to control the construction process and update predictions of movements at early stages of construction. In particular, any time a new set of monitoring displacements are available, the finite element model of the excavation system is "recalibrated" to provide the best fit to the field observations.

4.2 The case study

An inverse analysis procedure that uses construction monitoring data to update predictions of deformations for supported excavation systems is presented. The numerical procedure is used to optimize the finite element model of a deep excavation through Chicago glacial clays [Cal03, Cal04, Fin05]. The excavation consisted of removing 12.2 m of soft to medium clay within 2 m of a school supported on shallow foundations. The support system consisted of a secant pile wall supported by one level of cross-lot bracing and two levels of tie-backs. Ground movements during construction were recorded using inclinometers installed around the excavation site.

The commercial software PLAXIS 7.11 was used to compute the response of the soil around the excavation in plane-strain conditions (Figure 9). The Figure shows the central portion of the finite element mesh and the elevations at the interfaces of the different soil layers. The soil stratigraphy was assumed to be uniform across the site. Eight soil layers were modeled: a fill layer overlaying a clay crust, a compressible clay deposit consisting of four distinct clay layers, and a relatively incompressible deposit consisting of two clay layers. The fill layer was modeled as an elastic-perfectly plastic Mohr–Coulomb material, whereas all clays layers were modeled using the H-S model [Sch99]. This effective stress model is formulated within the framework of elastoplasticity; plastic strains are calculated assuming multisurface yield criteria; isotropic hardening is assumed for both shear and volumetric strains; the flow rule is nonassociative for frictional shear hardening and associative for the volumetric cap. The calculation phases used in the finite element simulations were 21, starting with the construction of both the tunnel tubes and the school adjacent to the excavation. Table 5 shows the five phases for which the model predictions are updated. Observations from two inclinometers on opposite sides of the excavation were used to compare computed displacements with the field data.

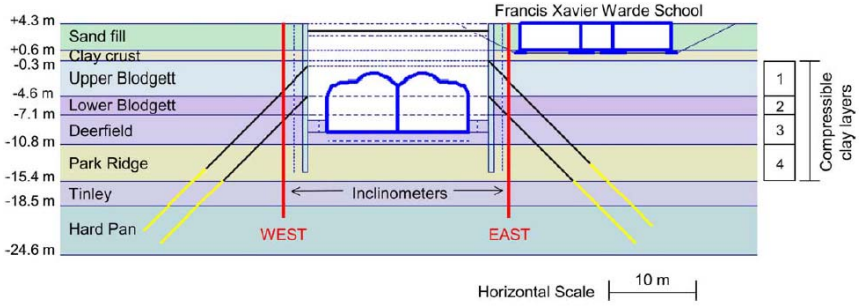


Figure 9: Schematic of numerical model of excavation system [Cal04].

Table 5: Excavation stages considered for updating model predictions.

		Construction day	
Stage	Plaxis phase	WEST	EAST
Initial	Beginning of project	0	0
1	Drill piles	11	18
2	Excavate and put struts	60	73
3	Excavate and prestress tiebacks (1st level)	88	95
4	Excavate and prestress tiebacks (2nd level)	105	109
5	Excavate to final depth [-7.9m]	112	123

Figure 10 shows the soil profile, a schematic of the support system and the observation points retrieved from the inclinometer data for the five construction stages considered. Inclinometer readings were taken in the field every two feet. Not every reading, however, could be used as an observation for the inverse analysis because the finite element displacements were computed only at the intersection between the finite element mesh and the inclinometer location. Thus, 13 observation points were used for the east side and 11 observation points for the west side. The observations for the last two stages on the west side are not reported because the inclinometer was destroyed by construction activities after Stage 3.

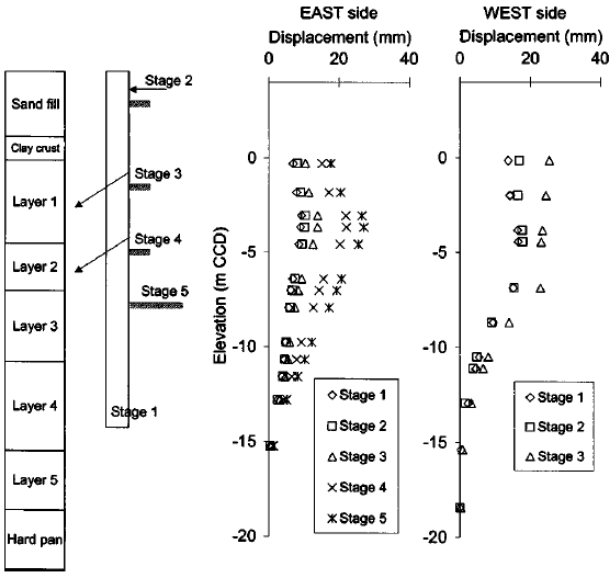


Figure 10: Schematic of retaining system and observations points used from inclinometer readings [Fin05].

Table 6 shows the initial values of the six basic H-S input parameters for the five clay layers that are calibrated by inverse analysis. These parameters are: friction angle, ϕ; cohesion, c; dilation angle, ψ; reference secant Young's modulus at 50% stress level, E_{50}^{ref}; reference oedometer tangent modulus, E_{oed}^{ref} ; exponent m. The latter parameter relates the reference moduli to the stress level dependent moduli E (E_{50}, E_{oed}, and E_{ur}):

$$E = E^{ref}\left(\frac{c\,cot\varphi - \sigma_3'}{c\,cot\varphi + p^{ref}}\right) \tag{17}$$

where: p^{ref} = reference pressure equal to 100 stress units; σ_3' = minor principal effective stress.

Layers 1 to 5 refer to the Upper Blodgett, Lower Blodgett, Deerfield, Park Ridge, and Tinley layers, respectively. The initial estimates of the input parameters for Layers 1 to 4 were based on triaxial test results. Because few laboratory data existed for the very stiff Layer 5 soil and very small movements were observed in that stratum, the initial values of the parameters for Layer 5 were selected to minimize movements in that stratum.

Table 6: Initial values of Hardening-Soil parameters for the clay layers.

Parameter	Layer 1	Layer 2	Layer 3	Layer 4	Layer 5
ϕ	23.4	23.4	25.6	32.8	32.8
c (kPa)	0.05	0.05	0.05	0.05	0.05
y	0	0	0	0	0
E_{50}^{ref} (kPa)	226	288	288	413	619
E_{oed}^{ref} (kPa)	158	202	202	289	433
m	0.8	0.8	0.85	0.85	0.85

4.3 Definition of the inverse analysis problem

The input parameters optimized by inverse analysis using UCODE were chosen following the procedure described in Fig. 8. The observations, soil movements, and the many types of loading paths associated with the excavation simulation are very different from the stress–strain data used as observations in the triaxial compression tests. The results of a sensitivity analysis performed on the H-S basic parameters indicated that the parameters that are most relevant to the excavation problem are E_{50}^{ref}, m and ϕ [Cal04]. Figure 11 shows the composite scaled sensitivities of the three relevant parameters for layers 1 to 5. The bar chart refers to sensitivities computed using all the observations, the line charts refer to sensitivities computed from the observations of the different layers. From a simulation perspective, results show that the parameters that most influence the simulation are the ones relative to layers 1, 3, and 4. Layer 1 is the softest soil layer, thus its major influence on the displacement results is expected. Layer 3 is the stratum wherein the excavation bottoms out. Layer 4 is the stiff clay layer below the bottom of the excavation into which the secant pile wall is tipped. The high sensitivity values of this stratum indicate that the strength and the stiffness of the clay below the excavation have significant impact on movements. The Figure also shows that the observations relative to a soil layer are mainly influenced by changes in that soil layer's parameters. Table 7 shows the correlation coefficients between the three parameters at every layer. The rather high correlation between E_{50}^{ref} and m indicate that these parameters are not likely to be simultaneously and uniquely optimized, even though the results of the analysis are sensitive to both. For calibration purposes parameter E_{50}^{ref} , rather than parameter m, was chosen to "represent" the stiffness of the H-S model because changes in E_{50}^{ref}

values also produce changes in the values of parameters E_{oed}^{ref} (equal to 0.7 times E_{50}^{ref}) and E_{ur}^{ref} (equal to 0.7 times E_{50}^{ref}), thus its calibration can be considered as representative of the calibration of all H-S stiffness parameters.

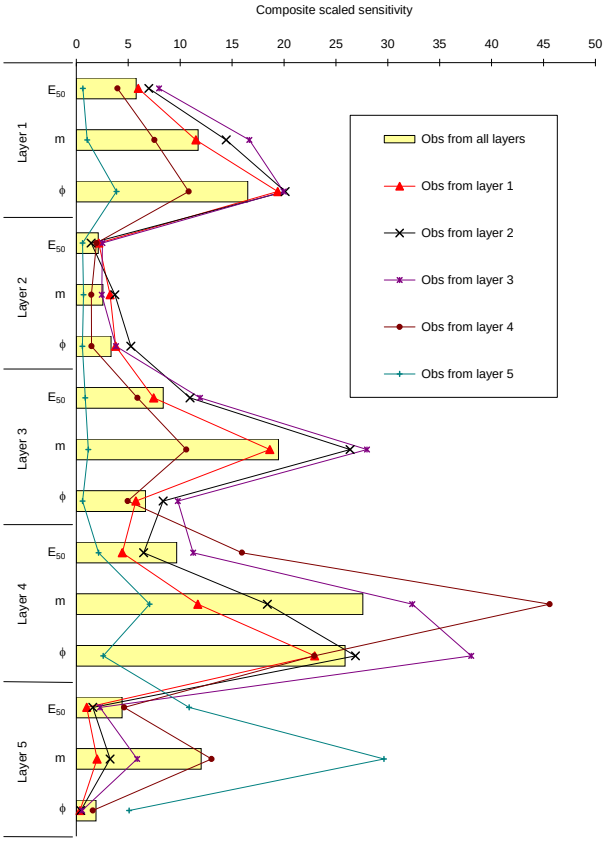


Figure 11: Composite scaled sensitivities of E^{ref}_{50}, m and ϕ for layers 1 to 5.

A final reduction of the parameters to optimize was necessary to establish a "well-posed" problem where the solution converged [Cal04]. To this aim, layers 1 and 2 were combined because: layer 2 had a much lower impact on the computed results, as indicated by the low values of composite scaled sensitivities; the two layers are derived from the same geologic stratum. Moreover, the stiffness parameters (E_{50}^{ref}) were chosen over the failure parameters (ϕ) because: the excavation-induced stress conditions in the soil around this excavation were, for the most part, far from failure; the laboratory estimated values of ϕ are judged to be more accurate than E_{50}^{ref}. Note that, when the stiffness and failure parameters are optimized simultaneously or only the failure parameters are calibrated, the regression analysis never converged to reasonable values. This emphasizes the point that convergence does not necessarily

ensure that reasonable results are attained when optimizing a nonlinear problem such as a supported excavation in soil. Table 8 shows the H-S parameters updated by the regression algorithm. Fifteen parameters were updated at every iteration, but only three of them (E1, E3, and E4) were directly estimated by the optimization algorithm. Note that changing the values of E_{50}^{ref} is not the same as merely changing the elastic parameters of an elastoplastic or linear elastic soil model because the hardening soil responses are nonlinear below the cap and the stiffness depends on more than E_{50}^{ref}. A summary of the inverse analysis setup is presented in Table 9.

Table 7: Correlation coefficients between E_{ref}^{50}, m and ϕ at every layer.

Layer	Between parameters	Value	Between parameters	Value	Between parameters	Value
1		-.70		-.42		.33
2	E_{ref}^{50}	-.85	E_{ref}^{50}	-.59	m	.41
3	and	-.87	and	-.58	and	.25
4	m	-.99	ϕ	-.07	ϕ	-.14
5		-.95		.39		-.56

Table 8: Parameters updated by inverse analysis.

Layer	Parameter optimized	Related parameter	
1	$E_1=E_{ref}^{50}(1)$	$E_{oed}^{50}(1)=0.7E_1$	$E_{ur}^{50}(1)=3E_1$
2	$E_2=E_{ref}^{50}(2)=E_1$	$E_{oed}^{50}(2)=0.7E_2$	$E_{ur}^{50}(2)=3E_2$
3	$E_3=E_{ref}^{50}(3)$	$E_{oed}^{50}(3)=0.7E_3$	$E_{ur}^{50}(3)=3E_3$
4	$E_4=E_{ref}^{50}(4)$	$E_{oed}^{50}(4)=0.7E_4$	$E_{ur}^{50}(4)=3E_4$
5	$E_5=E_{ref}^{50}(5)=1.5E_4$	$E_{oed}^{50}(5)=0.7E_5$	$E_{ur}^{50}(5)=3E_5$

Table 9: Summary of the inverse analysis setup.

Geotechnical variables	Observations	readings from inclinometers (west, east)
	Input parameters	1 parameter (E_{50}^{ref}) per layer (5 layers)
	Initial calibration	by inverse analysis from triaxial tests
	Discretization (west)	11 readings per construction stage
	Discretization (east)	13 readings per construction stage
Other variables	Observations' weighting	σ^2 = measurement error variance
	Convergence criteria	TOL = SOSR = 5%
	Regression variables	MAX-CHANGE = 0.5
	Sensitivity calculation	PERTURBATION = 0.01

4.4 Results

The simplest way to evaluate the difference between the results of the numerical simulations based on the initial estimates of the parameters and the optimized ones is to compare the inclinometer data with the computed horizontal displacements for the two cases. Figure 12 shows the visual fit between the observations and the results computed before (i.e., initial) and after (i.e., best-fit) the calibration by inverse analysis. The comparison shows that the initial simulation computes displacements significantly larger than the measured ones at every construction stage (the maximum computed displacements at stage 5 are about two times the measured ones) and the computed displacement profiles result in significant and unrealistic movements in the lower clay layers. When the model is calibrated by inverse analysis, the fit between the computed and measured response is extremely good. At the end of the construction the maximum computed displacement exceeds the measured data by less than 10% and the distributions of lateral deformations are consistent throughout the excavation. Note that the good fit shown in the Figure refers to the final optimization, i.e. all observations (stages 1-5) were used to calibrate the finite element model of the excavation. Yet, the simulation was calibrated starting at stage 1 and re-calibrated at every subsequent construction stage using the inclinometer data available up to that stage [Fin05].

Figure 13 shows the initial and the final values of error variance and objective function at each optimization stage. The graphs allow the comparison between the overall magnitude of weighted residuals relative to the initial estimates of the parameters and the fit resulting from the calibrated models. Results show that error variance values decrease by more than two orders of magnitude at every stage. In all cases, the final error variance values are close to 1.0, indicating that the computed differences are consistent with the measurement errors. More importantly, the results show that stage 1 observations improved the predictions by two orders of magnitude and that, by the end of stage 3, the recalibration of the model is essentially complete. In essence, they indicate that early observations are able to recalibrate the finite element simulation in a way that is beneficial to the predictions of movements at later stages. Note that, in this case, stage 1 refers to the installation of secant pile walls inducing movements throughout the compressible clay layers. A satisfactory calibration at this stage indicates that these movements were large enough to “exercise” the constitutive laws of all soil layers subsequently affected by the excavation. The variation of the input parameters at the five optimization stages is shown in Figure 14 above a bar chart, representing the progress of the excavation, in which the excavation depth is normalized with respect to the excavation width. Results show that the maximum changes in parameter values occur at Stage 1, when the observations relative to the installation of the secant pile wall are used. The results indicate that the initial estimates of the stiffness parameters are significantly lower than the optimized values of the parameters. This trend could be expected because the initial values were based on results of triaxial compression tests on specimens taken from thin-wall tubes. Yet, if an analyst was to arbitrarily increase the initial stiffness parameters prior to optimization, the magnitude of the increase would be a

matter of much judgment and, most likely, the parameter values would still require subsequent adjustments to provide good fits to the observed data.

The fact that initial parameters based on the laboratory data led to optimized parameters that resulted in a good fit, and made sense from a geotechnical viewpoint, illustrate the utility of the method. As implied in Figure 13, portions of the increase in optimized stiffness between stages 2 and 3 may be a result of end effects of the excavation. The simulated excavation is really a threedimensional problem modeled in plane strain. When the excavated depth is small, most of the wall can be adequately modeled as plane strain and, hence, little changes in parameters are noted between stage 1 and 2. As the excavation deepens, the ratio between excavation depth and excavation width increases and higher parameter values compensate for the lack of constraints in the out-of-plane direction.

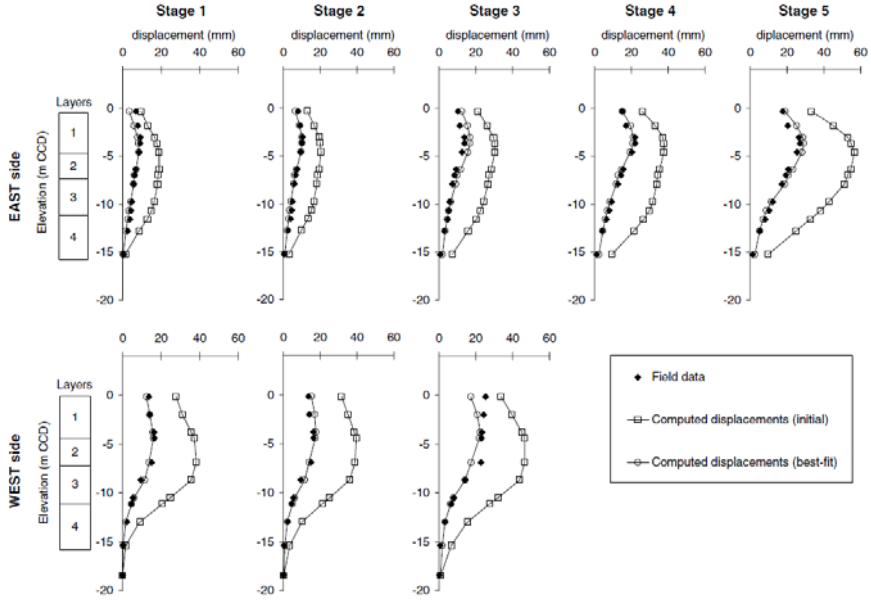


Figure 12: Measured and computed horizontal displacements for initial and best-fit estimates of parameters [Cal04].

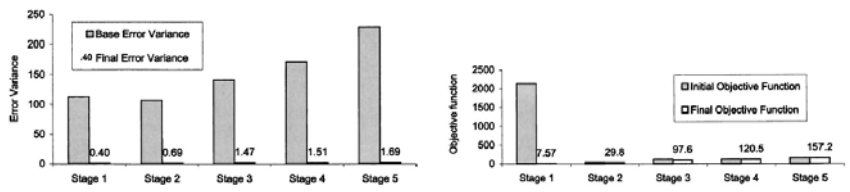


Figure 13: Error variance and objective function at each optimization stage [Fin05].

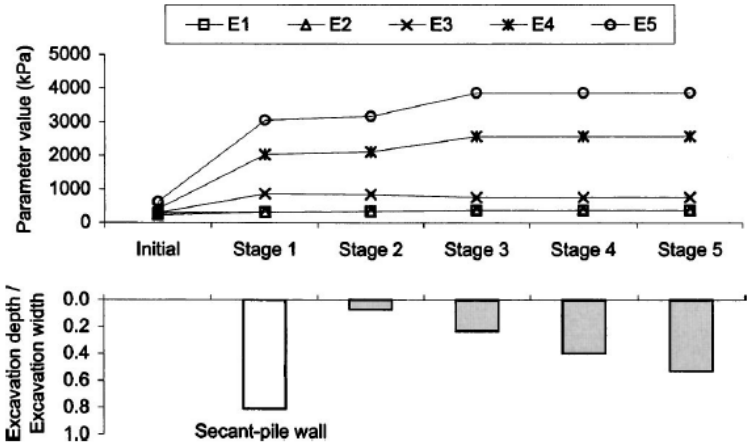


Figure 14: Best-fit parameter values and normalized excavation depth at different optimization stages [Fin05].

5 Conclusions

The inverse modeling procedures described in this work employ a parameter optimization algorithm (UCODE) to efficiently calibrate soil models by minimizing the errors between experimental observations and model response. The chapter showed how inverse analysis techniques could be effectively used for two very different geotechnical purposes: to calibrate the parameters of a soil constitutive law based on triaxial experimental results; to control the construction process of an excavation by updating the predictions of movements during construction based on the field monitoring data.

References

[Cal02] Calvello, M.. *Inverse analysis of a supported excavation through Chicago glacial clays*. PhD Thesis, Northwestern University, 2002.

[Cal03] Calvello, M., Finno, R.J.. Modeling excavations in urban areas: effects of past activities. *Rivista Italiana di Geotecnica, 37(4):9-23*, 2003.

[Cal04] Calvello, M., Finno, R.J.. Selecting parameters to optimize in model calibration by inverse analysis. *Computers and Geotechnics, 31(5):411-425*, 2004. DOI: 10.1016/j.compgeo.2004.03.004

[Fin91] Finno R.J. and Harahap I.S.. Finite element analysis for the HDR-4 excavation. *Journal of Geotechnical Engineering, ASCE, Vol. 117(8):1045-1064*, 1991.

[Fin02] Finno, R.J., Bryson, L.S., Calvello, M.. Performance of a stiff support system in soft clay. *Journal of Geotechnical and Environmental Engineering, ASCE, 128(8):660-671*, 2002. DOI: 10.1061/(ASCE)1090-0241(2002)128:8(660)

[Fin05] Finno, R.J., Calvello, M.. Supported excavations: observational method and inverse modeling. *Journal of Geotechnical and Environmental Engineering, ASCE, 131(7):826-836*, 2005. DOI: 10.1061/(ASCE)1090-0241(2005)131:7(826)

[Hil98] Hill M.C. (1998). *Methods and guidelines for effective model calibration*. U.S. Geological Survey Water-Resources investigations report 98-4005, 1998.

[Pec69] Peck R.B. (1969). Deep excavations and tunneling in soft ground. *Proc. 7th Int. Conf. on Soil Mechanics and Foundation Engineering, State-of-the-Art Volume: 225-290*, 1969.

[Poe98] Poeter E.P. and Hill M.C.. *Documentation of UCODE, a computer code for universal inverse modeling*. U.S. Geological Survey Water-Resources investigations report 98-4080, 1998.

[Ros58] Roscoe, K.H., Schoefied, A., Wroth C.P.. On yielding of soils. *Geotechnique, 8; 22–53*, 1958.

[Sch99] Schanz T., Vermeer P.A. and Bonnier P.G.. The Hardening Soil model - formulation and verification. *Proc. Plaxis Symp. Beyond 2000 in Computational Geotechnics, 281-296*, 1999.

Analysing time dependent problems

Cristina Jommi and Patrick Arnold

Delft University of Technology, The Netherlands

Inverse analysis for time dependent problems is discussed in this chapter. When time dependent processes are analysed, further uncertainties come from initial conditions as well as from time dependent boundary conditions and loads, in addition to model parameters. Inverse modelling techniques have been specifically developed for this class of problems, which exploit the availability of a set of measurement and/or monitoring data at given locations at subsequent time instants. Sequential Bayesian data assimilation is introduced, and a brief review of filtering techniques is given. In filtering the problem unknown is the time evolution of the probability density function of the system state, described by means of appropriate time dependent variables and time invariant parameters, conditioned to all previous observations. Particle filtering is chosen to conceptually illustrate the methodology, by means of two simple introductory examples.

1 Introduction

Many engineering systems, and most often geotechnical systems, show time dependent response, due to time dependent loads and boundary conditions, as well as to multiphysics coupling. To assess time dependent, *dynamic*, systems and to predict their evolution in time, a proper model to describe the behaviour of the soil has to be conceived and calibrated, the initial state has to be known, and the time evolution of boundary conditions and of loads has to be described. This adds further uncertainty to the comprehensive *model* we use to describe the physical system. Moreover, small scale laboratory tests can usually give only partial information on the material properties, which have to be upscaled to properly describe the response of the system at the field scale. This is true, in general, for any geotechnical property, but even more for the hydraulic behaviour of soils, which typically shows high non-linearities due to multiphysics coupling, and which is usually strongly affected by scale effects and heterogeneity. Without loss in generality, in this chapter reference is made to the hydraulic behaviour, although the derivations can be equally applied to any other

multiphysics process.

Iterative adjustment of the model parameters, including the soil property values, the initial conditions of the system, and the time dependent boundary conditions, is a powerful tool to infer the future state of a system, given the history of its observed previous response. *Sequential data assimilation* utilises inverse modelling to estimate the state of the dynamic system at each time a measurement or an observation from the system becomes available. In this context, we speak about *measurement* when a variable describing the state of the system is measured directly (e.g. pore water pressure), and about *observation*, when the state variable is inferred by measuring a related quantity, like water content or porosity from electro-magnetic sensors.

When looking at a general time dependent physical system, neither its *"real (true) state"*, $\hat{\mathbf{x}}$, nor the *"real (true) observation"*, $\hat{\mathbf{y}}$, at the current time, $t+1$, are known in reality. The true state is a function of the true history of the state $\hat{\mathbf{x}}|^{0\rightarrow t+1}$, the true and time invariant soil parameters $\hat{\mathbf{p}} = \{\hat{p}_1, \hat{p}_2, ...\}$ and the true history of the boundary conditions $\hat{\mathbf{u}}|^{0\rightarrow t+1}$, while the true observation is a function of the true state of the system at that time

$$\hat{\mathbf{x}}^{t+1}(\hat{\mathbf{x}}|^{0\rightarrow t+1}, \hat{\mathbf{p}}, \hat{\mathbf{u}}|^{0\rightarrow t+1}) \tag{1}$$

$$\hat{\mathbf{y}}^{t+1}(\hat{\mathbf{x}}^{t+1}) \tag{2}$$

with $\hat{\mathbf{x}} \in \mathcal{R}^{N_{\hat{x}}}$, where $N_{\hat{x}}$ is the dimension of the vector describing the system state at a given location, and $\hat{\mathbf{y}} \in \mathcal{R}^{N_{\hat{y}}}$ is the observation vector of dimension $N_{\hat{y}}$.

In order to analyse and infer the response of a time dependent systems, two types of models are required; (a) a model $\mathcal{M}$ of some form describing the transient processes affecting the state, and (b) a model $\mathcal{G}$ relating some observation of the processes to the system state.

(a) The state of a dynamic system is commonly assessed using a discrete-time approach. The *predicted state(s)*, $\mathbf{x}$, may be defined as a first order Markov process, that is, the system state at the current time, $t+1$, is only a function of the state at the previous time step t. Hence

$$\mathbf{x}^{t+1} = \mathcal{M}(\mathbf{x}^t, \mathbf{p}, \mathbf{u}^t) + \epsilon_{\mathbf{x}}^{t+1} \tag{3}$$

where $\mathcal{M}$ is the model operator describing the non-linear physical process as a function of the state $\mathbf{x}^t$ at time t, the time invariant model parameters $\mathbf{p} \in \mathcal{R}^{N_p}$ and the prescribed model boundary conditions $\mathbf{u}^t$ at time t. The Gaussian (white) noise term $\epsilon_{\mathbf{x}} \sim \mathcal{N}(0, \sigma^2_{\epsilon_{\mathbf{x}}})$ (see Chapter 1, [Fen14]) is adding stochastic diffusion and has a mean of zero and a variance of $\sigma^2_{\epsilon_{\mathbf{x}}}$.

The state can be also written as an augmented state variable

$$\mathbf{z}^{t+1} = (\mathbf{x}^{t+1}, \mathbf{p}^{t+1}) \tag{4}$$

[e.g. RHV10, MDS12]. As the parameters in Equations 3 are time invariant, this augmented state variable may be used to describe the estimation of the parameters $\mathbf{p}$ with respect to the state at time $t+1$.

(b) The observation at $t+1$ can be computed by

$$\mathbf{y}^{t+1} = \mathcal{G}(\mathbf{x}^{t+1}, \mathbf{p}) + \epsilon_{\mathbf{y}}^{t+1} \tag{5}$$

where $\mathcal{G}$ is the measurement function of the system response and $\epsilon_{\mathbf{y}} \sim \mathcal{N}(0, \sigma^2_{\epsilon_{\mathbf{y}}})$ is the Gaussian noise term of the observation.

For most time dependent non-linear soil processes, the inference of the state, as well as of the soil property values, from direct inversion using closed-from analytical frameworks is virtually impossible, as already discussed in the previous chapters [Led14, Cal14]. The local gradient-based search algorithms, previously introduced to iteratively determine the local minimum within a maximum likelihood and weighted least square framework, become more likely to fail in finding the global minimum with increasing non-linearity of the system. Indeed, these algorithms are not designed to handle highly multivariate problems with multiple local optima in parameter space and multiple domains of attraction, and they become less and less effective with increasing domain size or in the presence of discontinuous responses [VSW$^+$08]. More robust global optimisation algorithms have been developed, that use multiple searches from different starting points within the parameter space, to reduce the risk of attraction towards a single local domain. The classical inference methods for non-linear dynamic systems are the *Kalman filters* and its variants, which have been have been successfully applied to many non-linear problems. An alternative sequential Monte Carlo method, the *particle filter*, has been chosen here to introduce the potentials and the limitations of global optimisation methods for time dependent processes. After a brief general introduction to sequential Bayesian data assimilation, a review of sequential inference is given. The Particle Filter is then briefly illustrated, and discussed by means of two introductory examples.

2 Inverse modelling

2.1 Bayesian basics

A background on Bayesian theory is provided in Chapter 1 [Fen14] or in specific monographs [e.g. BT92, Gre05] and thus will be summarised here only briefly. The basic form of Bayes' theorem in a continuous version is

$$P[E_i|A] = \frac{P[A|E_i]P[E_i]}{P[A]} \tag{6}$$

where E_i is the event, i.e. the state to be predicted, A is the occurrence, i.e. the observed data/measurments, $P[E_i]$ is the prior distribution or expectation defining the prior knowledge of the event i ($\int P[E_i] = 1$), $P[A|E_i] \propto P[E_i|A]P[A]$ is the likelihood function ($\int P[A|E] \neq 1$), $P[E_i|A]$ is the posterior distribution estimating

the event E_i given the observed data A ($\int P[E|A] = 1$) and $P[A]$ represents the marginal distribution of A ($\int P[A] = 1$), i.e. the normalisation factor

$$\begin{aligned} P[A] &= P[A|E]P[E] + P[E|A^c]P[A^c] \\ &= P\left[\bigcup_{i=1}^{n}(A \cap E_i)\right] = \sum_{i=1}^{n} P\left[A \cap E_i\right] = \sum_{i=1}^{n} P[A|E_i]P[E_i] \end{aligned} \tag{7}$$

Inserting Equation 7 into Equation 6 the posterior distribution can be written as

$$P[E_i|A] = \frac{P[A|E_i]P[E_i]}{\sum_{i=1}^{n} P[A|E_i]P[E_i]} \tag{8}$$

A simple example will illustrate how a Bayesian scheme can be applied. Let us assume that the volumetric water content θ has to be inferred to describe the state of an unsaturated soil, and that, based on previous experience, laboratory tests, or database, we know that $\theta \sim \mathcal{N}(\mu_\theta, \sigma_\theta^2)$ with a variance of $\sigma_\theta^2 = 0.0009$. A set of new direct laboratory measurements of θ from soil samples retrieved in the field becomes available, which allows updating our prior knowledge on the mean μ_θ. Given the prior of the mean $\mu_\theta = \Theta \sim \mathcal{N}(\mu_\Theta, \sigma_\Theta^2)$

$$f(\Theta) = \sqrt{2\pi\sigma_\Theta^2}\ \exp\left\{-\frac{(\mu_\Theta - \Theta)^2}{2\sigma_\Theta^2}\right\} \tag{9}$$

the likelihood of the mean given one measurement θ is proportional to the likelihood of the sample mean $\bar{\theta}$ for a set of N_s independent measurements.

$$p(\theta|\Theta) = \sqrt{2\pi\sigma_\theta^2}\ \exp\left\{-\frac{(\theta - \Theta)^2}{2\sigma_\theta^2}\right\} \propto \sqrt[N_s]{2\pi\sigma_\theta^2}\ \exp\left\{-\frac{\left(\bar{\theta} - \Theta\right)^2}{2\sigma_\theta^2/N_s}\right\} \tag{10}$$

Therefore, the likelihood is normally distributed with a mean Θ, a variance σ_θ^2/N_s and the shape being controlled by the sample size. The posterior distribution is then obtained via multiplication of the prior and likelihood function

$$p(\Theta|\theta) \propto \exp\left\{-\frac{\left(\bar{\theta} - \Theta\right)^2}{2\sigma_\theta^2/N_s} - \frac{(\mu_\Theta - \Theta)^2}{2\sigma_\Theta^2}\right\} \tag{11}$$

where $p(\Theta|\theta) \sim \mathcal{N}(\tilde{\mu}, \tilde{\sigma})$ with a mean $\tilde{\mu} = \frac{\Theta\sigma_\theta^2/N_s + \sigma_\Theta^2\bar{\theta}}{\sigma_\Theta^2 + \sigma_\theta^2/N_s}$ and variance $\tilde{\sigma} = \frac{\sigma_\Theta^2\sigma_\theta^2/N_s}{\sigma_\Theta^2 + \sigma_\theta^2/N_s}$.

It's worth noting that the example can be extended to the case of unknown mean and variance, given that the state variable can be described by a normal or log-normal distribution.

The quality of the prior information and the advantage provided by the new measurements are illustrated in Figure 1, where N_s is the number of samples available. Given $f(\Theta) \sim \mathcal{N}(0.42, 0.0009)$ and $p(\theta|\Theta) \sim \mathcal{N}(0.33, 0.18^2/N_s)$ taking only three samples the likelihood is low, and the prior distribution dominates the posterior (Figure

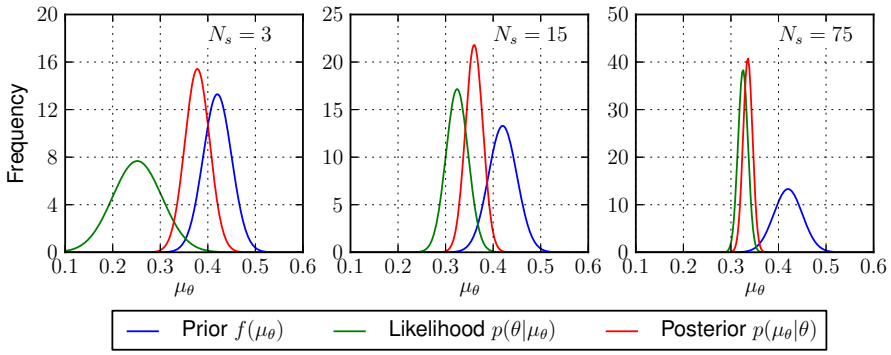


Figure 1: Example of Bayesian inference of normally distributed mean estimate μ_θ of a sample θ.

1(a)). With increasing sample size, the confidence in the observation increases and the prior information becomes less significant for the posterior prediction (Figure 1(b-c)), The difference between the prior and the posterior distributions indicates that the prior information was poor in this case.

Two simplifying assumptions were implicitly introduced in the previous derivation: (i) that the water content does not change in time, and (ii) that it can be sampled by means of direct measurement. In a more realistic scenario, (i) the water content in the field will be time dependent, as it is the result of soil-atmosphere interaction and of the position of the groundwater table, and (ii) non-invasive observations, often based on electromagnetic techniques, are usually preferred to track the physical process. Therefore, the information available usually refers to a sequential distribution of states, and a model, including its own uncertainty (Equation 5), has to be introduced to translate the observations into water content sampling. To deal with information referring to a time dependent sequence of states, the Bayesian scheme can be extended over time.

2.2 Sequential Bayesian Data Assimilation

Given the initial joint *probability density function* (PDF) of the state $p(\mathbf{x}_0|\mathbf{y}_0) \equiv f(\mathbf{x}_0)$, where $\mathbf{y}_0$ indicates no observation, the aim of this assimilation process is to sequentially infer the posterior state $p(\mathbf{x}^{0:t+1}|\mathbf{y}^{1:t+1})$ at present time conditioned by any observation which became available in time. This characterisation of the distribution state of the hidden Markov tracking process is referred to as *filtering* [DJ11].

Given the state $E = \mathbf{z}$ (Equation 4), and the observation $A = \mathbf{y}$ (Equation 5), Equation 8 gives the conditional posterior PDF at time $t + 1$

$$p_{\mathbf{Z}|\mathbf{Y}}(\mathbf{z}^{0:t+1}|\mathbf{y}^{1:t+1}) = \frac{p_{\mathbf{Y}|\mathbf{Z}}(\mathbf{y}^{1:t+1}|\mathbf{z}^{0:t+1})p_{\mathbf{Z}}(\mathbf{z}^{0:t+1})}{p_{\mathbf{Y}}(\mathbf{y}^{1:t+1})} \tag{12}$$

where $p_{\mathbf{Y}|\mathbf{Z}}(\mathbf{y}^{1:t+1}|\mathbf{z}^{0:t+1})$ is the likelihood, $p_{\mathbf{Z}}(\mathbf{z}^{0:t+1})$ is the prior and $p_{\mathbf{Y}}(\mathbf{y}^{1:t+1})$ is scaling the numerator to satisfy $\int p_{\mathbf{Z}|\mathbf{Y}}(\mathbf{z}^{0:t+1}|\mathbf{y}^{1:t+1}) = 1$. For simplicity the subscripts $\mathbf{Z}$ and $\mathbf{Y}$, indicating the random nature of $\mathbf{z}$ and $\mathbf{y}$, will be omitted from here on.

The above posterior distribution (Equation 12) joins information of all past states and is commonly referred to as *optimal filtering problem* [e.g. DJ11]. Recursion of Equation 12 satisfying the marginal filtering posterior distribution $p(\mathbf{z}^{t+1}|\mathbf{y}^{1:t+1})$, holding only information of the current state, can be written as

$$p(\mathbf{z}^{t+1}|\mathbf{y}^{1:t+1}) = \frac{g(\mathbf{y}^{t+1}|\mathbf{z}^{t+1})p(\mathbf{z}^{t+1}|\mathbf{y}^{1:t})}{p(\mathbf{y}^{t+1}|\mathbf{y}^{1:t})} \tag{13}$$

where g is the homogeneous likelihood (state and observation densities are time independent) and the prior distribution is estimated as

$$p(\mathbf{z}^{t+1}|\mathbf{y}^{1:t}) = \int_{\mathbf{z}^t} f(\mathbf{z}^{t+1}|\mathbf{z}^t, \mathbf{y}^{1:t})p(\mathbf{z}^t|\mathbf{y}^{1:t})\ \mathrm{d}\mathbf{z}^t = \int_{\mathbf{z}^t} f(\mathbf{z}^{t+1}|\mathbf{z}^t)p(\mathbf{z}^t|\mathbf{y}^{1:t})\ \mathrm{d}\mathbf{z}^t \tag{14}$$

also known as *Chapman-Kolmogorov equation*, which simplifies to the second term due to the first order Markov process [DJ11, MDS12]. The likelihood g is commonly described by a Gaussian with zero mean and a given variance. The normalisation factor may be predicted using the augmented state as intermediate variables [MDS12].

$$p(\mathbf{z}^{t+1}|\mathbf{y}^{1:t}) = \int_{\mathbf{z}^{t+1}} g(\mathbf{y}^{t+1}|\mathbf{z}^{t+1})p(\mathbf{z}^{t+1}|\mathbf{y}^{1:t})\ \mathrm{d}\mathbf{z}^{t+1} \tag{15}$$

Equation 15 is commonly referred to as the *predictive/evolution* step and Equations 12 with 14 are referred to as the *updating/correction* step [e.g. DdG01, CGM07, DJ11].

2.3 Sequential inference of soil water dynamic processes

In geotechnical engineering, the use of inverse models in sequential schemes is still lagging behind. Explicit analysis of transient processes by using advanced constitutive models implemented in numerical frameworks is usually preferred, also due to the difficulties in obtaining a comprehensive body of statistically valuable in situ measurements. The state and the parameters of transient processes have been inferred by using gradient based optimization algorithms, with their application ranging, for example, from deep staged excavations [e.g. RLF08, TK09] to laboratory pulse test [e.g.

GGA+11]. However, in these schemes the different sources of uncertainty are not accounted for, and no or only limited information on the state and parameters are being carried from one measurement time step to the next one. The Bayesian framework presented in Section 2.2 is not exploited in these cases.

Sequential data assimilation is a very commonly applied tool, for instance, in weather forecasting, hydrological modelling and flood protection assessment. For most soil water dynamic processes the non-linearity in the soil response and transient boundary conditions, alongside with the non-Gaussianity of the distributions, makes an analytical solution of the Equations 13-15 untraceable [e.g. CGM07, PCP12]. To overcome this limitation, [Eve94] developed a recursive data-processing algorithm known as the *ensemble Kalman filter* (EnKF), which is an extension to the original *Kalman filter* [Kal60] and the *extended Kalman filter* [Jaz70]. The EnKF is based on a *Markov chain Monte Carlo* (MCMC) method, propagating a large ensemble of model states to approximate the prior state error in time by using the updated states of the previous time step, to predict the current ensemble via forward integration of a stochastic differential equation describing the model dynamics [e.g. BvE98, Eve03, Eve09]. More information, examples and codes can be found on Geir Evensen's EnKF-homepage[1].

The EnKF is one of the most commonly used non-linear filter for state and parameters updating in many fields such as hydrological modelling [e.g. MSGH05, PCP12, SNH12], but it has not often been adopted in the assimilation of geotechnical systems [e.g. HMHV10, CCZ10].

The *particle filter* (PF), which is a *sequential Monte Carlo* (SMC) method, presents one alternative to the EnKF. The PF method is very flexible, easily implementable, strongly parallelisable and, most importantly, it approximates the probability densities directly via a finite number of samples, often referred to as *particles* [DdG01, AMTC02]. A large number of different PF methods was developed in recent years. Some tutorials and state-of-the-art reports provide a good introduction and allow for a more complete overview [e.g. DdG01, DJ11, LW01, AMTC02, CGM07, van09, CR11]. Some useful resources on SMC and PF methods have been compiled by Arnaud Doucet[2].

Most PF frameworks are based on a *sequential importance sampling* (SIS) and *sampling importance resampling* (SIR) algorithm. The SIS is the most basic Monte Carlo method to approximate the prediction and the updating steps (Equations 12-15). It uses a finite set of random samples with associated weights to directly represent the posterior distribution at current time step, and subsequently updates this particles in order to obtain the posterior at the next time step. However, for non-linear systems the sample may tend to degenerate, that is, only a limited number of particles being around the *"real"* state exclusively carry the weights, whilst the remaining majority of samples only carry a negligible weight. To increase the effectiveness of the filter and avoid errors accumulation, the SIR algorithms may be used. SIR introduces a resampling stage at each time step, in which particles with a low weight are eliminated

[1] EnKF sources: `http://enkf.nersc.no/`
[2] SMC and PF sources: `http://www.stats.ox.ac.uk/~doucet/smc_resources.html`

and regenerated in zones in which particles carry a high weight, which renders the approximation of the posterior. Other PF methods include auxiliary particle filters, marginalised particle filters, Markov chain particle filters and may incorporate some particle smoothing algorithm [e.g. AMTC02, CGM07, DJ11]. Sequential smoothing makes use of the estimates of the past states and thus tends to provide a better filter for the current state.

In recent years the PF method became popular and performed well in the assimilation of the state and parameters of different hydrological soil water dynamic processes [e.g. MHGS05, MDS12, KdD05, SF09, QLY$^+$09, RHV10, MMW$^+$11, NTSK11, PDD$^+$12, RVS$^+$12]. A comparison between the EnKF and PF performance using a coupled surface-subsurface flow model has been presented by [PCP12].

In geotechnical engineering the use of the PF method is not common. However, Murakami and co-workers [SMN$^+$12, MSN$^+$13] recently demonstrated that the elastic-plastic Cam Clay model parameters can be successfully inferred using a coupled hydro-mechanical Finite Element program in a PF framework, both on synthetic observation data for soil element loading tests and the construction of a soil embankment, as well as on real observation data related to the construction of the Kobe Airport Island.

3 A simple SIR particle filter implementation

The posterior (Equation 13) is approximated using a discrete set of N_s samples

$$p(\mathbf{x}^{0:t+1}|\mathbf{y}^{1:t+1}) = \sum_{k=1}^{N_s} w_k^{t+1} \delta\left(\hat{\mathbf{x}}^{t+1} - \mathbf{x}_k^{t+1}\right) \tag{16}$$

where w_k^{t+1} are the normalised particle weights

$$w_k^{t+1} = \frac{w_{k^*}^{t+1}}{\sum_{k=1}^{N_s} w_{k^*}^{t+1}} \tag{17}$$

When using the transient prior as importance function, i.e. $q(\mathbf{x}_k^{t+1}|\mathbf{x}_k^t, \mathbf{y}^{t+1}) = p(\mathbf{x}^{t+1}|\mathbf{x}_k^t)$, the updated sequential estimates of the importance weights are

$$w_{k^*}^{t+1} \propto w_{k^*}^t \; \frac{p(\mathbf{y}^{t+1}|\mathbf{x}_k^{t+1}) p(\mathbf{x}^{t+1}|\mathbf{x}_k^t)}{q(\mathbf{x}_k^{t+1}|\mathbf{x}_k^t, \mathbf{y}^{t+1})} = w_{k^*}^t \; p(\mathbf{y}^{t+1}|\mathbf{x}_k^{t+1}) \tag{18}$$

which represent the key part of the SIS filter [e.g. AMTC02, MHGS05, DJ11].

The implementation of a simple SIR filter based on [MHGS05] is schematised in Figure 2. The process can be split into three stages.

Initialisation stage:

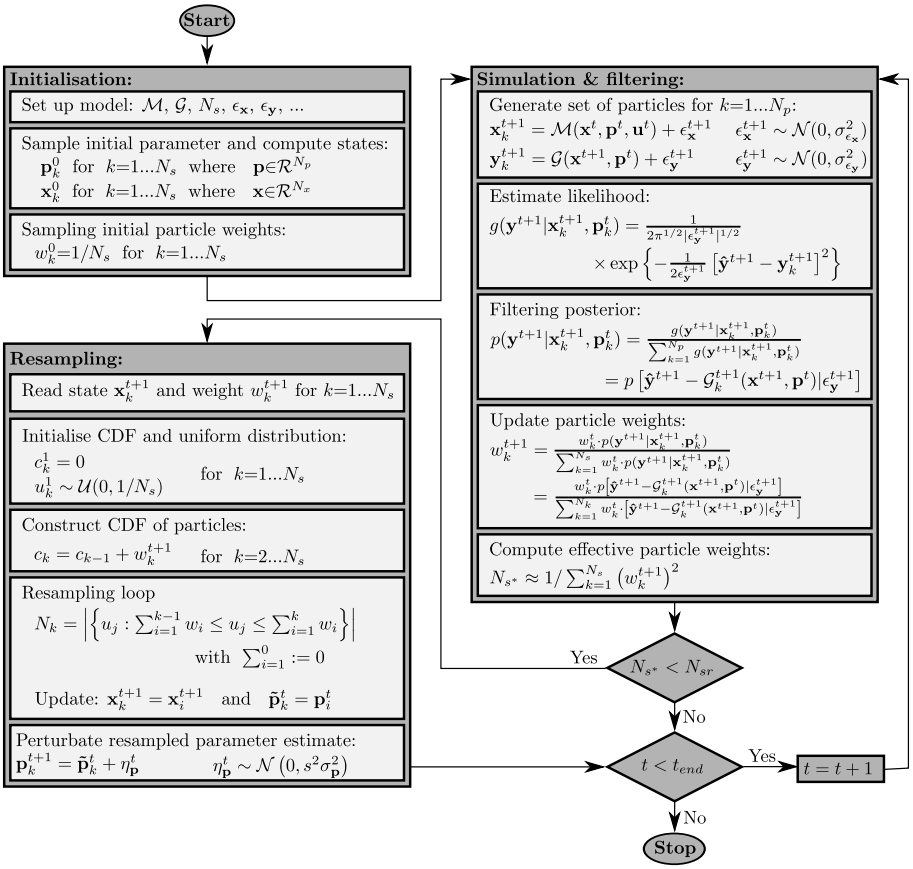


Figure 2: Schematic description of a simple SIR PF.

In this first stage the process and the observation model, $\mathcal{M}$ and $\mathcal{G}$, (Equations 3 and 5) as well as the stochastic model, e.g. the number of particles (samples) N_s and the error functions ϵ, are set up. The initial state $\mathbf{x}^0$ is computed based on a set of parameters $\mathbf{p}^0$ representing the prior knowledge, and an initial set of uniform weights w_k^0 is assigned to each particle k.

Simulation stage:

In the simulation stage, the filtering of the state at $t + 1$ is performed. By means of the state and the observation models, $\mathbf{x}^{t+1}$ and $\mathbf{y}^{t+1}$ are computed for each particle. Subsequently the homogeneous likelihood function g is estimated to compute the filtering posterior (Equations 13 and 16). Utilising Equations 17 and 18 the weights are assigned to each the particle. Given that the effective particle size N_{s^*} is smaller than a minimum effective particle size N_{sr}, representing a resampling threshold below

which degeneration of the samples occurs, the resampling stage is entered.

Resampling stage:

Different schemes for state and parameters resampling have been proposed. Using one of the systematic schemes [e.g. AMTC02, DJ11], the particles are resampled by relating a *cumulative distribution function* (CDF) for the particle, c, to a uniform CDF, u. After the update of the particle states and parameters, the resampled parameter estimate $\tilde{\mathbf{p}}_k^t$ is perturbed to obtain

$$\mathbf{p}_k^{t+1} = \tilde{\mathbf{p}}_k^t + \eta_{\mathbf{p}}^t \tag{19}$$

with $\eta_{\mathbf{p}}^t \sim \mathcal{N}(0, s^2\sigma_{\mathbf{p}}^2)$ being a Gaussian noise term, as suggested by [LW01] and [MHGS05]. The variance of the parameter particles, $\sigma_{\mathbf{p}}^2$, is multiplied by a small tuning parameter s, which determines the exploration radius around each particle, and for which values between 0.005 and 0.025 have been commonly used [MDS12].

4 Examples

Two introductory examples will be discussed in this section to demonstrate the working principle and the efficiency of the simple SIR PF implementation. In the first benchmark example synthetic observations are used, while the second example refers to a typical field case where direct measurements are available.

4.1 Example 1: an analytical benchmark

The first example has been used as benchmark as well as for illustrative purpose by several authors [e.g. KdD05, MHGS05]. The non-linear state model and the observation function are both one-dimensional and described by the following analytical functions

$$x^{t+1} = \frac{1}{2}x^t + a\frac{x^t}{1+(x^t)^2} + b\cos(1.2t) + \epsilon_x \tag{20}$$

$$y^{t+1} = \frac{(x^{t+1})^2}{20} + \epsilon_y \tag{21}$$

where $a = 25$ and $b = 8$ are the *parameters*, $\epsilon_x \sim \mathcal{N}(0, \sigma_{\epsilon_x}^2)$ and $\epsilon_y \sim \mathcal{N}(0, \sigma_{\epsilon_y}^2)$ are the random noise terms for the state and the observation respectively, with $\sigma_{\epsilon_x}^2 = 10$ and $\sigma_{\epsilon_y}^2 = 1$. The initial state is taken as $x^0 = 10$ and $N_s = 500$ particles are used. The initial parameter estimates are $a^0 = 30$ and $b^0 = 4$.

Figure 3 shows the state, the observation and the inferred parametric response with time, using a sampling time interval of $\Delta t = 1$. The results of the simulation show

that the sequential assimilation technique succeeds in predicting quickly and accurately the state, with some negative peaks not being detected until the end of the simulation. During the time lapse analysed, also the parameters converge close to the real values ($m_{a^{100}} = 25.1$ and $m_{b^{100}} = 7.869$). The remaining variation of the parameters depends partly on the tuning parameter s, which ensures that the filter were able to react to any significant variation in the observation.

In this benchmark case, resampling was required in most of the time steps to avoid degeneracy. Figure 4 exemplary illustrates the resampling scheme for the last time step $t = 100$. The reduction of the variance of the filtering posterior from (b) to (d) due to resampling allows for a more effective use of the particles. A more detailed description of the filtering and resampling process can be found for instance in [NTSK11], [MDS12] and [SMN$^+$12].

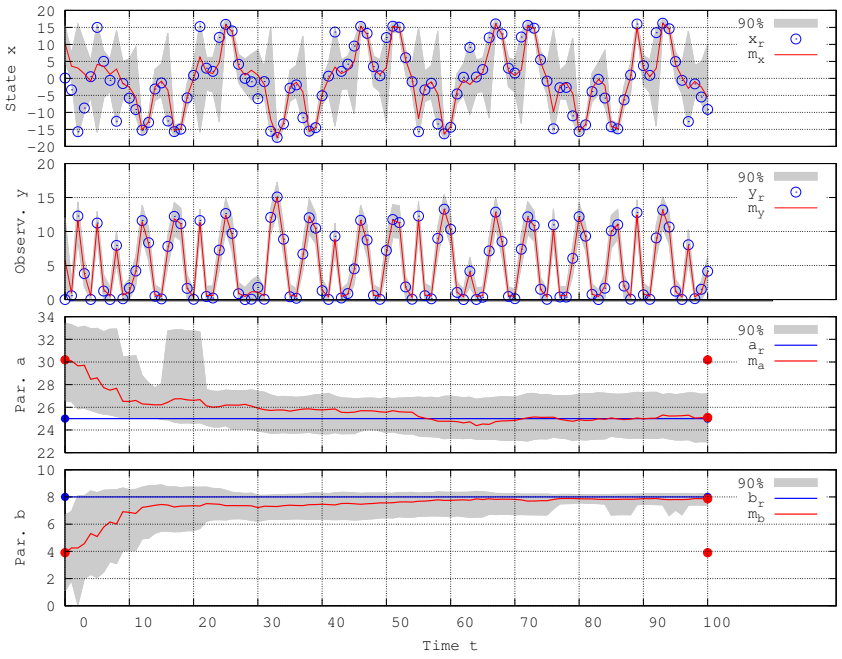


Figure 3: Sequential assimilation of state and parameters using a SIR PF with 500 particles showing the state x and observation y response as well as the evolution of the parameters a and b with time t. Subscripts r indicate the synthetic *"real"* values, m the sample mean and 90% the 90% bounds, respectively.

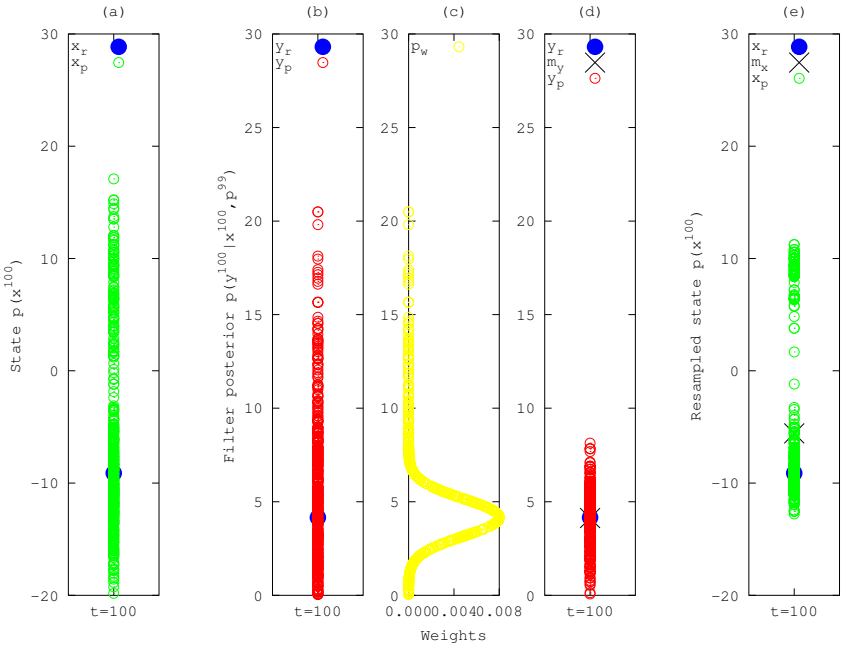


Figure 4: Particle resampling at time step $t = 100$ showing (a) the state, (b) the filter posterior, (c) the weights assigned to each particle and (d-e) the resampled posterior and state. Subscripts r indicate the synthetic *"real"* values, subscripts p the particles and m the sample mean, respectively.

4.2 Example 2: response of pore water pressure below a dike

In the assessment of dikes, different failure mechanisms have to be analysed. Most of them are likely to be initiated by the transient pore groundwater response to the time dependent external forcing conditions. The worst conditions are not necessarily associated to the steady state pore water distribution in equilibrium with the maximum expected water height. Therefore, proper assessment of the potential failure mechanisms requires the analysis of the fully coupled time dependent hydro-mechanical response of the water defense structure, including the dike body and the subsoil.

Explicit coupled numerical finite elements analyses can be performed to this aim, but the computational effort needed to include uncertainty in the model is still high. A valuable alternative consists in relying on simplified analytical solutions of the coupled hydro-mechanical consolidation process, and perform an inverse analysis able to sequentially assimilate the parameters of the simplified model by comparison with observation in time. Figure 5 gives a simplified illustration of the hydro-mechanical

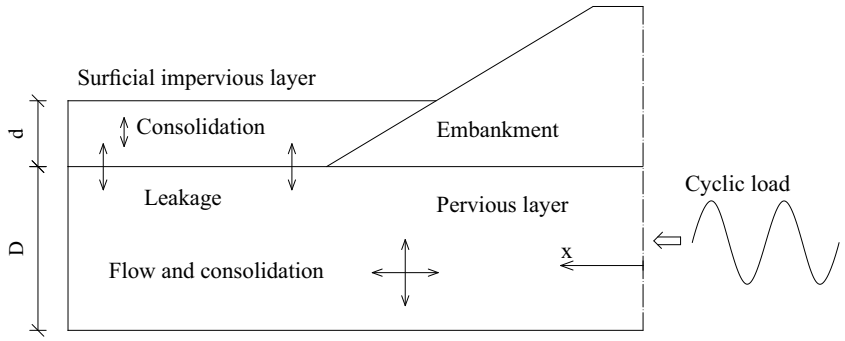


Figure 5: Simplified description of the hydro-mechanical response of a dike subsoil subjected to cyclic hydraulic boundary conditions

Table 1: Parameters of the simplified consolidation model.

Variable		Unit	$\hat{\mathbf{p}}$	$\mathbf{p}^0$
Thickness of clay layer	d_c	[m]	2.0	–
Thickness of sand layer	D_s	[m]	7.0	6.25
Sat. hydr. conductivity of clay layer	K_c	[m s^{-1}]	0.00001	–
Sat. hydr. conductivity of sand layer	K_s	[m s^{-1}]	0.0005	0.005
Compressibility of clay layer	α_c	[m kN^{-2}]	0.005	–
Compressibility of sand layer	α_s	[m kN^{-2}]	0.0000001	–

processes taking place in the typical foundation subsoil of a dike, with a pervious aquifer underlying an impervious surficial layer, subjected to a cyclic variation of the pore water pressure at the boundary representing the river bed. The effective simple analytical solution proposed by Baudin & Barends [BB88] for this problem has been used in this second example.

The adopted analytical solution gives the pore pressure distribution in the two layers at any given distance from the river bed, and is a function of the variables listed in Table 1. The response of the system depends on the thickness, on the compressibility and on the hydraulic conductivity of the two layers, and on the period of the forcing boundary condition. Table 1 summarises the synthetic soil property values $\hat{\mathbf{p}}$ representing the *real* state of the system $\hat{\mathbf{x}}$ (Equation 1) assumed to be represented by the analytical solution. Prior investigation using Monte Carlo simulations had shown that the response of the hydraulic head $H(x)$ in the sand layer is most sensitive to variations in the saturated hydraulic conductivity of the sand layer K_s, while the thickness of the sand layer D_s has a less dominating role. For the sake of simplicity, in this example the random model parameters are limited to this two variables, for which the initial guess is $\mathbf{p}^0 = \{D_s, K_s\} = \{6.25, 0.005\}$, and the remaining four parameters are assumed to be known.

We assumed - as this is the case in the field test to which this example refers to - that a piezometer is installed in the sand layer at a distance $x = 18.6$m from the river bed, where a direct measurement of the pore water pressure is taken at each hour. These pore pressure measurements are used for sequential data assimilation, using the SIR PF previously described.

The period of the forcing function is $T = 10$d (86 400s), and the measurements are taken at each hour (3 600s). The total time was set to 200h (720 000s), the number of particles to $N_s = 400$ and s to 0.005. The maximum hydraulic head at the river was normalised to $h_0(x = 0) = 1.0$m.

Figure 6 shows the parameter estimate with time. The hydraulic conductivity converges very rapidly to the synthetic *"real"* value, confirming the high potential of the adopted algorithm in sequential data assimilation. On the contrary, the convergence for the thickness of the sand layer is much slower, and the uncertainty does not decrease monotonically. Indeed, the two variables were chosen with the purpose of assessing the performance of the algorithm in identifying parameters which have different relative weights on the prediction. The hydraulic conductivity of the pervious layer, which dominates the response of the system, could be rapidly inferred. As for the thickness of the same layer, a reasonable convergence could be achieved, in spite of its minor role in the response of the synthetic *true* system.

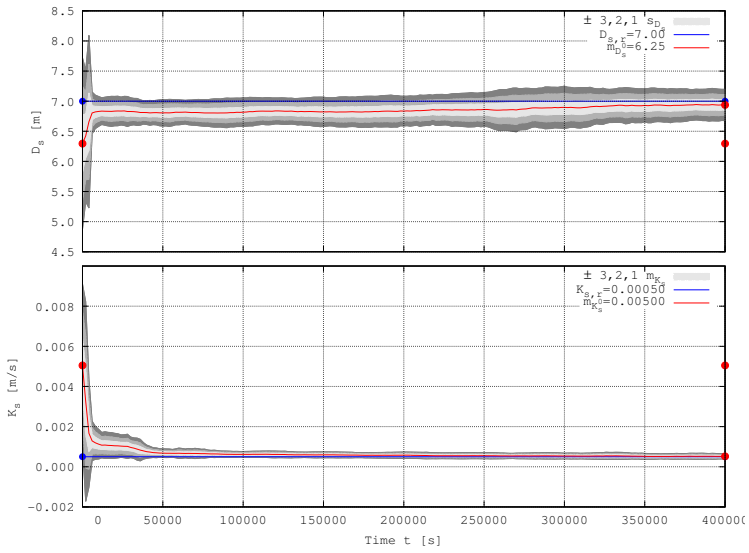


Figure 6: Estimation of the thickness of sand layer, D_s, and the saturated hydraulic conductivity of the sand layer, K_s, using a SIR PF. The subscripts r indicate the synthetic *real* soil property values, the superscript 0 the initial state, and m and s the sample mean and standard deviation, respectively.

5 Final remarks

In this last chapter, a basic introduction to the inverse analysis of time dependent problems was given. The provided overview is far from being a complete review, and to this aim the reader is referred to the references for further reading.

The two basic aims of this contribution were: (i) to combine the theoretical and numerical developments on random fields, presented in the first part of this book, with the general concepts on inverse analysis illustrated in the previous two chapters; and (ii) to open a window on sequential data assimilation, which can be fruitfully exploited in the practice, when time dependent problems have to be analysed. The examples discussed at the end of this chapter are meant just as an introduction to the powerful approaches which can be adopted in these cases. Nonetheless, they suggest that if information from measurement and monitoring in *space* of a time dependent system is accompanied by a thorough analysis of the observed behaviour in *time*, identification of the variables and parameters dominating the response of the models can be effectively accomplished by means of rather simple dedicated algorithms.

References

[AMTC02] M. S. Arulampalam, S. Maskell, N. Gordon T., and Clapp. A tutorial on particle filters for on-line non-linear/non-Gaussian Bayesian tracking. *Transactions on Signal Processing*, 50(2):174–188, 2002.

[BB88] C. M. H. L. G. Baudin and F .B. J. Barends. Getijderespons in grondwater onder Nederlandes dijken. *H2O*, 21(1):2–5, 1988.

[BT92] G. E. P. Box and G. C. Tiao. *Bayesian Inference in Statistical Analysis*. Wiley Classics Library. John Wiley & Sons, Inc., New York, US, 1992.

[BvE98] G. Burgers, P. J. van Leeuwen, and G. Evensen. Analysis Scheme in the Ensemble Kalman Filter. *Monthly Weather Review*, 126(6):1719–1724, 1998.

[Cal14] M. Calvello. *Calibration of soil constitutive laws by inverse analysis*, chapter Review of Probability Theory. ALERT Doctoral School. ALERT Geomaterials, 2014.

[CCZ10] H. Chang, Y. Chen, and D. Zhang. Data assimilation of coupled fluid flow and geomechanics using the ensemble Kalman filter. *SPE Journal*, 15(2):382–394, 2010.

[CGM07] O. Cappé, S. J. Godsill, and E. Moulines. An overview of existing methods and recent advances in sequential Monte Carlo. *Proceedings of the IEEE*, 95(5):899–924, 2007.

[CR11] D. Crisan and B. Rozovskii. *The Oxford handbook of nonlinear filtering*. Oxford University Press, Oxford, UK, 2011.

[DdG01] A. Doucet, N. de Freitas, and N. Gordon. *Sequential Monte Carlo Methods in Practice*, chapter An Introduction to Sequential Monte Carlo Methods, pages 3–14. Statistics for Engineering and Information Science. Springer, 2001.

[DJ11] A. Doucet and A.M. Johansen. *The Oxford handbook of nonlinear filtering*, chapter A Tutorial on Particle Filtering and Smoothing: Fifteen years Later. Oxford University Press, Oxford, UK, 2011.

[Eve94] G. Evensen. Sequential data assimilation with a non-linear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Research: Oceans (1978–2012)*, 99(C5):10143–10162, 1994.

[Eve03] G. Evensen. The Ensemble Kalman Filter: theoretical formulation and practical implementation. *Ocean Dynamics*, 53(4):343–367, 2003.

[Eve09] G. Evensen. *Data Assimilation: The Ensemble Kalman Filter*. Springer, 2 edition, 2009.

[Fen14] G. A. Fenton. *Stochastic Analysis and Inverse Modelling*, chapter Review of Probability Theory. ALERT Doctoral School. ALERT Geomaterials, 2014.

[GGA+11] R. Giot, A. Giraud, C. Auvray, F. Homand, and T. Guillon. Fully coupled poromechanical back analysis of the pulse test by inverse method. *International Journal for Numerical and Analytical Methods in Geomechanics*, 35(3):329–359, 2011.

[Gre05] P. C. Gregory. *Bayesian Logical Data Analysis for the Physical Sciences: A Comparative Approach with Mathematica Support*. Cambridge University Press, New York, US, 2005.

[HMHV10] A . Hommels, F . Molenkamp, M . Huber, and P . A . Vermeer. Inverse modelling including spatial variability applied to the construction of a road embankment. In *Numerical Methods in Geotechnical Engineering (NUMGE 2010)*, pages 369–374. CRC Press, 2010.

[Jaz70] A. H. Jazwinski. *Stochastic Processes and Filtering Theory*. Academic Press, New York, 1970.

[Kal60] R. E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering (ASME)*, 82(Series D):35–45, 1960.

[KdD05] M. Klaas, N. de Freitas, and A. Doucet. Toward Practical N^2 Monte Carlo: the Marginal Particle Filter. In F. Bacchus and T. Jaakkola, editors, *Proceedings of the Twenty-First Conference Conference on Uncertainty in Artificial Intelligence (UAI-05)*, Edinburgh, Scotland, 6 2005.

[Led14] A. Ledesma. *Geotechnical back-analysis using a maximum likelihood approach*, chapter Review of Probability Theory. ALERT Doctoral School. ALERT Geomaterials, 2014.

[LW01] J. Liu and M. West. *Sequential Monte Carlo Methods in Practice*, chapter Combined Parameter and State Estimation in Simulation-Based Filtering, pages 197–223. Statistics for Engineering and Information Science. Springer, 2001.

[MDS12] H. Moradkhani, C. M. DeChant, and S. Sorooshian. Evolution of ensemble data assimilation for uncertainty quantification using the particle filter-Markov chain Monte Carlo method. *Water Resources Research*, 48(12):W12520, 2012.

[MHGS05] H. Moradkhani, K.-L. Hsu, H. Gupta, and S. Sorooshian. Uncertainty assessment of hydrologic model states and parameters: Sequential data assimilation using the particle filter. *Water Resources Research*, 41(5):W05012, 2005.

[MMW+11] C. Montzka, H. Moradkhani, L. Weihermüller, H.-J. H. Franssen, M. Canty, and H. Vereecken. Hydraulic parameter estimation by remotely-sensed top soil moisture observations with the particle filter. *Journal of Hydrology*, 399(3–4):410–421, 2011.

[MSGH05] H. Moradkhani, S. Sorooshian, H. V. Gupta, and P. R. Houser. Dual state-parameter estimation of hydrological models using ensemble kalman filter. *Advances in Water Resources*, 28(2):135–147, 2005.

[MSN+13] A. Murakami, T. Shuku, S. Nishimura, K. Fujisawa, and K. Nakamura. Data assimilation using the particle filter for identifying the elasto-plastic material properties of geomaterials. *International Journal for Numerical and Analytical Methods in Geomechanics*, 37(11):1642–1669, 2013.

[NTSK11] S. J. Noh, Y. Tachikawa, M. Shiiba, and S. Kim. Applying sequential Monte Carlo methods into a distributed hydrologic model: lagged particle filtering approach with regularization. *Hydrology and Earth System Sciences*, 15:3237–3251, 2011.

[PCP12] D. Pasetto, M. Camporese, and M. Putti. Ensemble kalman filter versus particle filter for a physically-based coupled surface-subsurface model. *Advances in Water Resources*, 47:1–13, 2012.

[PDD+12] D. A. Plaza, R. De Keyser, G. J. M. De Lannoy, L. Giustarini, P. Matgen, and V. R. N. Pauwels. The importance of parameter resampling for soil moisture data assimilation into hydrologic models using the particle filter. *Hydrology and Earth System Sciences*, 16:375–390, 2012.

[QLY+09] J. Qin, S. Liang, K. Yang, I. Kaihotsu, R. Liu, and T. Koike. Simultaneous estimation of both soil moisture and model parameters using particle

filtering method through the assimilation of microwave signal. *Journal of Geophysical Research: Atmospheres*, 114(D15103):13, 2009.

[RHV10] J. Rings, J. A. Huisman, and H. Vereecken. Coupled hydrogeophysical parameter estimation using a sequential Bayesian approach. *Hydrology and Earth System Sciences*, 14:545–556, 2010.

[RLF08] C. Rechea, S. Levasseur, and R. Finno. Inverse analysis techniques for parameter identification in simulation of excavation support systems. *Computers and Geotechnics*, 35(3):331–345, 2008.

[RVS+12] J. Rings, J. A. Vrugt, G. Schoups, J. A. Huisman, and H. Vereecken. Bayesian model averaging using particle filtering and Gaussian mixture modeling: Theory, concepts, and simulation experiments. *Water Resources Research*, 48(W05520):20, 2012.

[SF09] P. Salamon and L. Feyen. Assessing parameter, precipitation, and predictive uncertainty in a distributed hydrological model using sequential data assimilation with the particle filter. *Journal of Hydrology*, 376(3–4):428–442, 2009.

[SMN+12] T. Shuku, A. Murakami, S. Nishimura, K. Fujisawa, and K. Nakamura. Parameter identification for Cam Clay model in partial loading model tests using the particle filter. *Soils and Foundations*, 52(2):279–298, 2012.

[SNH12] A. Schöniger, W. Nowak, and H.-J. Hendricks Franssen. Parameter estimation by ensemble Kalman filters with transformed data: Approach and application to hydraulic tomography. *Water Resources Research*, 48(W04502):18, 2012.

[TK09] Y. Tang and G. T. Kung. Application of nonlinear optimization technique to back analyses of deep excavation. *Computers and Geotechnics*, 36(1–2):276–290, 2009.

[van09] P. J. van Leeuwen. Particle filtering in geophysical systems. *Monthly Weather Review*, 137(12):4089–4114, 2009.

[VSW+08] J. A. Vrugt, P. H. Stauffer, Th. Wöhling, B. A. Robinson, and V. V. Vesselinov. Inverse modeling of subsurface flow and transport properties: A review with new developments. *Vadose Zone Journal*, 7(2):843–864, 2008.

INPG – 3SR
46 avenue Félix Viallet
BP 53
38041 GRENOBLE CEDEX 9
FRANCE

ISBN 978-2-9542517-5-2

Fon: +33 (0) 456 528 621
Fax: +33 (0) 476 827 043
president@alertgeomaterials.eu
http://alertgeomaterials.eu

ALERT Doctoral School 2014

Stochastic Analysis and Inverse Modelling

Editors: M. A. Hicks, C. Jommi

ISBN 978-2-9542517-5-2