Probabilistic Modelling. An example-based guide

Book · April 2020

CITATION
1

READS
82

1 author:

L. Augusto Sanabria
Independent researcher (formerly with Geoscience Australia now retired)
75 PUBLICATIONS **927** CITATIONS

SEE PROFILE

Probabilistic Modelling.
An example-based guide

L.A. Sanabria

Formerly with the Risk and Impact Analysis Group - RIAG

Geoscience Australia

Version 1.1.
April 8, 2020

Prologue

There are very good books about mathematical modelling but there are very few books dealing with the specific subject of probabilistic modelling. The few books on this subject matter are however mostly theoretical in nature with little information about practical implementation of the techniques presented. That is the gap the present book tries to fill out.

Mathematical modelling is both an art and a science and the only way for the student to develop the skills necessary to become a fluent practitioner is by looking at what other people have been doing.

Probabilistic modelling is a subset of mathematical modelling, in the former the variables considered are random variables. These types of variables allow the modeller to include the uncertainty inherent in a large number of problems. In general, when we have to deal with future behaviour of a system we need to capture the possible range of variation of the system parameters and the most likely values those parameters can take. This can only be accomplished by using probabilistic modelling.

This book is for postgraduate students or practitioners who want to develop probabilistic models in their work. A basic knowledge of The Mathematical Theory of Probability is assumed. Something to the level of Gnedenko (1978) or Meyer (1972) will be adequate to understand most of the topics presented in this book.

As the book's title indicates the theory will be presented in the context of solving the problem discussed in the chapter. Enough details and formal proof of the basic principles will be given to show the reader the importance of sound mathematical principles for the solution of practical problems.

The basic technique to solve equations involving random variables is the so-called method of cumulants. This tecnique is not only mathematically robust but also very effective for the solution of these types of problems. It is also easy to implement in small computers and its efficiency allows the analyst to solve even large-scale problems from their desktop computer.

Chapter 1 introduces the basic concepts of statistical moments and cumulants and presents the basic algebra of cumulants which will be used for solution of linear equations. A few simple examples are presented all along the chapter. Chapter 2 presents a couple of realistic sized example problems taken from the field of electrical power systems. These examples show the tremendous power of the method of cumulants to solve probabilistic problems.

This book will be written chapter by chapter and new versions of the book will be uploaded to the internet to allow colleagues interested in these topics to try out some of the techniques discussed here. I will be grateful for comments from readers of this book, especially if they find errors. Please send your comments to: LASF1327@gmail.com.

Chapter 1

Solution of linear equations involving random variables using the method of cumulants (1)

1.1 Aim

The basic concepts on probabilistic modelling using the method of cumulants will be introduced in this chapter. These concepts will be used in subsequent chapters.

1.2 Introduction

Linear models are fundamental tools to solve a large type of problems. For this reason computer scientists have developed very efficient routines to solve these types of problems. Linear models are also widely used in probabilistic problems, however there not seems to be corresponding software for solution of linear models involving random variables. One of the aims of this book is to present efficient methods for solution of probabilistic models which can easily be implemented in a computer. This chapter starts by presenting some basic definitions. Proof of the rules to solve probabilistic linear equations is presented in the corresponding Section.

1.3 Methodology

The problem at hand is to solve a linear equation involving random variables (rvs). We start by looking at linear equations of independent random variables, the method will be extended to the case of rvs with some degree of correlation (non-independent rvs) later.[1]

Consider the equation,

$$Z = aX + bY + c \tag{1.1}$$

where,

a,b,c = constants
X,Y = Independent random variables
Z = Resulting random variable

The convolution of rvs X and Y to produce rv Z can be a very easy procedure if they are Normal distributions, however limiting the modeling of uncertainties to only Normal distributions is inappropriate in most modeling problems, so a general method to do the convolution of X and Y should be found. In most cases an analytical solution is impossible or very difficult to find, so alternative methods for solution of the general convolution problem have been proposed, one such method is the method of Cumulants.

In this method the rvs are replaced by a set of parametric values called Cumulants and using algebra, the Cumulants of the unknown rv Z can be calculated. The process does not involve any approximation. The limitation with this method is that the distribution of the unknown rv Z has to be calculated from its Cumulants, this can be achieved only through approximated methods as it will be discussed later.

[1] **Citation:** Sanabria L.A. (2018). Chapter 1: Solution of linear equations involving random variables using the method of cumulants. In: Probabilistic Modelling. An example-based guide (book in prep.)

1.4 Definitions

Cumulants are linear combinations of central moments. Central moments (CM) are moments with respect to the Mean of the distribution, they are given by the expression,

$$\mu_r = \mathbf{E}\{[X - \mathbf{E}\{X\}]^r\} \tag{1.2}$$

where,

μ_r = CM of order 'r' of the rv X
E{.} = Expected Value of a rv

Recall the expression for the calculation of the first moment (or *mean*) of rv X

$$\begin{aligned} m_1 = \mathbf{E}\{X\} = \sum_{\forall i} x_i \, p_i \qquad & \text{for the case of discrete distributions} \\ and, & \\ m_1 = \mathbf{E}\{X\} = \int_x x f(x) dx \qquad & \text{for the case of continuous distributions} \end{aligned} \tag{1.3}$$

in general the r-th *raw* moment of a random variable can be calculated using the familiar expressions,

$$m_r = \sum_{\forall i} x_i^r \, p_i \tag{1.4}$$

$$m_r = \int_x x^r f(x) dx \tag{1.5}$$

for discrete and continuous rvs respectively

Kendall & Stuart (1950) list all CM in terms of moments up to order 4:

$$\begin{aligned} \mu_1 &= 0 \\ \mu_2 &= m_2 - m_1^2 \\ \mu_3 &= m_3 - 3m_1 m_2 + 2m_2^3 \\ \mu_4 &= m_4 - 4m_1 m_3 + 6m_1^2 m2 - 3m_1^4 \end{aligned} \tag{1.6}$$

the book also lists cumulants from CM:

$$\begin{aligned} \kappa_1 &= \mu_1 \\ \kappa_2 &= \mu_2 \\ \kappa_3 &= \mu_3 \\ \kappa_4 &= \mu_4 - 3\mu_2^2 \\ \kappa_5 &= \mu_5 - 10\mu_3\mu_2 \end{aligned} \tag{1.7}$$

a practical way to calculate cumulants of any order from CM is given by the expression (Breitenberger, 1991),

$$\kappa_{r+1} = \mu_{r+1} - \sum_{j=1}^{r} A(r,j)\mu_j \kappa_{r-j+1} \tag{1.8}$$

where,

$$\text{A(r,j)} = \frac{r!}{j!(r-j)!}$$

In the next Section we will develop the algebra of cumulants, a fundamental tool to solve eq. of random variables. To facilitate the work we will use the convention:

$$\kappa_1 = m_1 \tag{1.9}$$

in other words, we define the first cumulant as equal to the mean of the distribution.

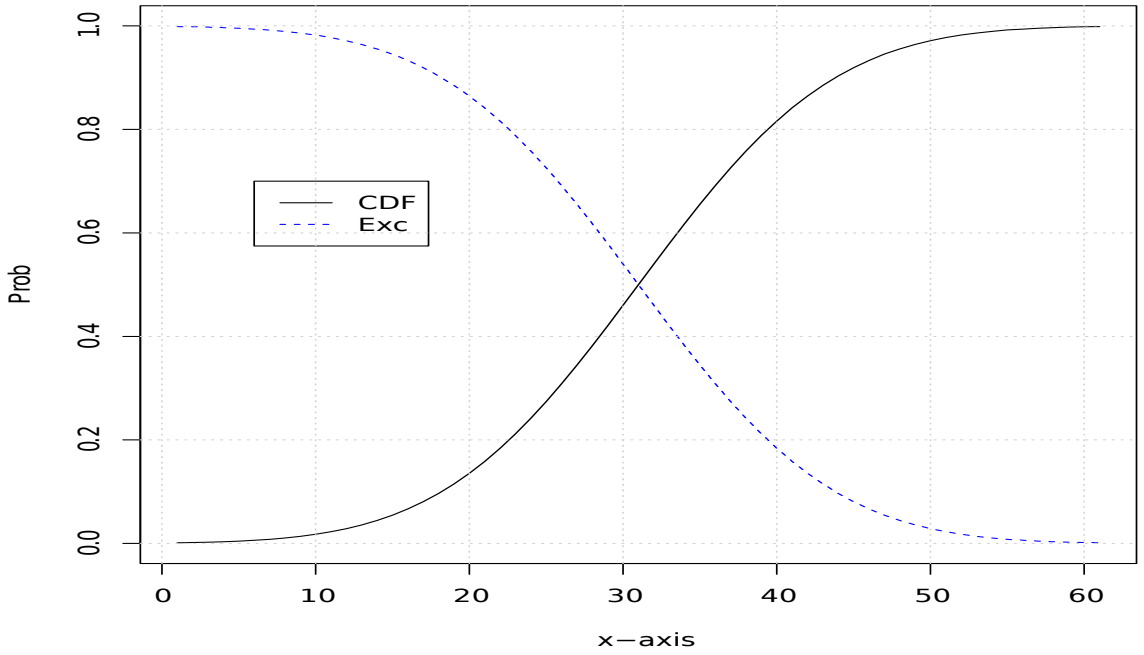

Figure 1.1: Cumulative and Exceedance distributions

In this and subsequent chapters we will make repetitive use of two well-known functions in probabilistic modelling: the first one is the Cumulative Distribution Function, F(x); the other one is the Exceedance Distribution, Exc(x), which is the complementary of F(x). The Cumulative Distribution Function or simply the distribution funcion of a continuous rv is given by,

$$F(x) = Pr[X \leq x] = \int_{-\infty}^{x} f(s)ds \tag{1.10}$$

Where "f(s)" is the density function of the rv. The corresponding Exceedance distribution is given by,

$$Exc(x) = Pr[X > x] = 1.0 - F(x) = \int_{x}^{\infty} f(s)ds \tag{1.11}$$

F(x) gives the probability that a rv X takes on a value less than or equal to x. Exc(x) gives the probability that the value x is exceeded. Fig. 1.1 shows both distributions on the same graph.

1.5 Algebra of Cumulants

The cumulants of the resulting rv Z (see eq. 1.1) can be calculated using expression,

$$\begin{aligned} \kappa_1(z) &= a\kappa_1(x) + b\kappa_1(y) + c \quad \text{first cumulant} \\ \kappa_2(z) &= a^2\kappa_2(x) + b^2\kappa_2(y) \quad \text{second cumulant} \\ \kappa_r(z) &= a^r\kappa_r(x) + b^r\kappa_r(y) \quad \text{cumulant of order } r > 2 \end{aligned} \tag{1.12}$$

To prove these expressions we need a couple more definitions, let's start with Characteristic functions (Lukacs, 1960) . The Characteristic function f_p(t) of a rv P which has F_p(x) as its distribution function (See Fig. 1.1), is given by the following definition,

$$f_p(t) = E[exp(itP)] = \int_{-\infty}^{\infty} e^{itx} dF_p(x) \tag{1.13}$$

where,

E[.] = mathematical expectation operator
i = imaginary operator ($\sqrt{(-1)}$)

The other definition we need is given by Thiele (1931), it is the Cumulant Generating function,

$$log(f_p(t)) = \sum_{r=1}^{s} \frac{\kappa_r^{(P)}}{r!}(it)^P + o(t^s) \tag{1.14}$$

where,

$\kappa_r^{(P)}$ = r-th order cumulant of rv P
log = log_e
o(t^s) is the error of the expansion

We want to prove that the convolution or combination of independent rvs can be expressed as a sum of their cumulants. Consider the simpler expression,

$$Z = X + Y + W + \ldots \tag{1.15}$$

The Characteristic function of Z can be written as,

$$f_z(t) = f_x(t) f_y(t) f_w(t) \ldots \tag{1.16}$$

$$f_z(t) = E[e^{it[X+Y+W+\ldots]}] = E[e^{itX} e^{itY} e^{iyW} \ldots] = E[e^{itX}]E[e^{itY}]E[e^{itW}] \ldots \tag{1.17}$$

and taking log of $f_z(t)$ we have,

$$log(f_z(t)) = log(f_x(t)) + log(f_y(t)) + log(f_w(t)) + \ldots \tag{1.18}$$

The proof of eq. 1.15 is completed by replacing $log(f_x(t))$ by eq. 1.14 and collecting terms of equal powers. ■

If the expression involves constants the solution is only a bit more complicated. Consider the expression,

$$\theta = aP + b \tag{1.19}$$

where,

a,b = constants
P = rv

$f_\theta(t)$, the Characteristic function of the rv θ which has distribution function $F_\theta(x)$, is given by,

$$f_\theta(t) = e^{itb} f_p(at) \tag{1.20}$$

because,

$$f_\theta(t) = \int_{-\infty}^{\infty} e^{it[ax+b]} dF_p(x) = e^{itb} \int_{-\infty}^{\infty} e^{itax} dF_p(x) = e^{itb} f_p(at) \tag{1.21}$$

To relate the cumulants of θ, κ_r, to the cumulants of the summands, eq. 1.14 should be used,

$$log(f_\theta(t)) = log[e^{itb} f_p(at)] = \sum_{r=1}^{s} \frac{\kappa_r}{r!}(it)^r + o(t^s) \tag{1.22}$$

that is,

$$(itb) + log(f_p(at)) = \sum_{r=1}^{s} \frac{\kappa_r}{r!}(it)^r + o(t^s) \tag{1.23}$$

hence,

$$(itb) + \sum_{r=1}^{s} \frac{\kappa_r^{(P)}}{r!}(iat)^r + o(at^s) = \sum_{r=1}^{s} \frac{\kappa_r}{r!}(it)^r + o(t^s) \tag{1.24}$$

For r = 1, we have,

$$(itb) + \kappa_1^{(P)}(iat) = \kappa_1(it) \tag{1.25}$$

hence,

$$\kappa_1 = a\kappa_1^{(P)} + b \tag{1.26}$$

For $r > 1$ we expand the summation and collect terms of equal powers,

$$(iat)^r \frac{\kappa_r^{(P)}}{r!} = (it)^r \frac{\kappa_r}{r!} \tag{1.27}$$

to obtain,

$$\kappa_r = a^r \kappa_r^{(P)} \tag{1.28}$$

■

Although we dont know the distribution of rv Z, knowledge of its cumulants gives us a lot of information about the rv. In some applications knowledge of the cumulants is enough to solve a problem. In particular, notice that κ_1 is the mean of the distribution. κ_2 is the variance (square of standard deviation) κ_3 is proportional to its skewness (asymmetry). κ_4 is proportional to its kurtosis (how peaky the distribution is). Notice also that in most cases the distribution is determined uniquely by its moments (Kendall & Stuart, 1950).

There are a number of methods to calculate the distribution of rv Z from its parameters, either its cumulants or its moments. Let us look at some of them.

1.6 Calculation of the pdf of rv Z from its parameters

1.6.1 Case A: We know that rv Z is a discrete distribution

. In this case we can use one of two methods: Von Mises Step Function or Hasofer three-point equivalent distribution. The later is a subset of the former.

Von Mises Step Function (VMSF):.

This method tries to find an r-step function which has the same 2r-1 known moments. If the set of known parameters is, in fact, a set of moments of a distribution, the method finds this distribution (Von Mises, 1964). So the first issue is to find whether a distribution exists which has those same moments or a subset of them. The solution of this problem is based on the theory of quadratic forms. The idea is to find the determinants D of the following arrays:

$$D_0 = \left|m_0\right| \tag{1.29}$$

$$D_1 = \begin{vmatrix} m_0 & m_1 \\ m_1 & m_2 \end{vmatrix} \tag{1.30}$$

$\vdots$

$$D_{r-1} = \begin{vmatrix} m_0 & m_1 & \cdots & m_{r-1} \\ \vdots & \vdots & \vdots & \\ m_{r-1} & m_r & \cdots & m_{2r-2} \end{vmatrix} \tag{1.31}$$

where,

$$m_r = \text{moment of order 'r'}$$
$$m_0 = 1$$

If all these determinants are positive, there exists a distribution which has those 2r-1 moments. Otherwise the moments of the last positive determinant are the ones which determine the distribution. This test is fundamental because it gives the number of impulses the discrete distribution has. Next, using the set of moments of the last determinant of (12), it is possible to find the r x-values in which the impulses are located. The x-axis values are the roots a_1, a_2, $a_3 \cdots a_r$ of the polynomial,

$$x^r + c_{r-1}x^{r-1} + c_{r-2}x^{r-2} + \cdots \; c_1 x + c_0 \tag{1.32}$$

The coefficients c_j are the solution of the system of linear equations,

$$[\mathrm{D}_{r-1}][c] = [m]$$

where $[\mathrm{D}_{r-1}]$ comes from the last positive determinant and

$$[\mathrm{m}] = [\text{-}\mathrm{m}_r - m_{r+1} \cdots \; - m_{2r-1}]^{\,T}$$

The probabilities of the discrete distribution (sometimes called impulses) corresponding to these x-values are calculated from the system of linear equations given by the Vandermonde matrix,

$$\begin{bmatrix} a_1^0 & a_2^0 & a_3^0 & \cdots & a_r^0 \\ a_1^1 & a_2^1 & a_3^1 & \cdots & a_r^1 \\ a_1^2 & a_2^2 & a_3^2 & \cdots & a_r^2 \\ \vdots & \vdots & \vdots & & \\ a_1^{r-1} & a_2^{r-1} & a_3^{r-1} & \cdots & a_r^{r-1} \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ \vdots \\ p_r \end{bmatrix} = \begin{bmatrix} m_0 \\ m_1 \\ m_2 \\ \vdots \\ m_{r-1} \end{bmatrix} \tag{1.33}$$

Three-point eqquivalent distribution:. Hasofer three-point equivalent distribution is much simpler to implement. The problem can be formulated as: Given a continuous distribution find its equivalent three-point (discrete) distribution (Hasofer, 2007).

The first step is to calculate the Mean and the first four CM of the continuous distribution applying expression (5). Suppose that the calculated Mean and CM are:

$$\begin{aligned} m_1 = &\quad \text{mean} \\ \mu_2 = &\quad \text{central moment of order 2} \\ \mu_3 = &\quad \text{central moment of order 3} \\ \mu_4 = &\quad \text{central moment of order 4} \end{aligned} \tag{1.34}$$

The equivalent three-point distribution can be calculated from these values as follows:

point	x-value	probability
1	$m_1 - y$	q
2	m_1	$1.0 - p - q$
3	$m_1 + x$	p

(1.35)

where,

$$\begin{aligned} x &= \Delta + \mu_3/2\mu_2 \\ y &= \Delta - \mu_3/2\mu_2 \\ \Delta &= \sqrt{(4\mu_4\mu_2 - 3\mu_3^2)}/2\mu_2 \end{aligned} \tag{1.36}$$

and,

$$\begin{aligned} p &= \mu_2/(x(x+y)) \\ q &= \mu_2/(y(x+y)) \end{aligned} \tag{1.37}$$

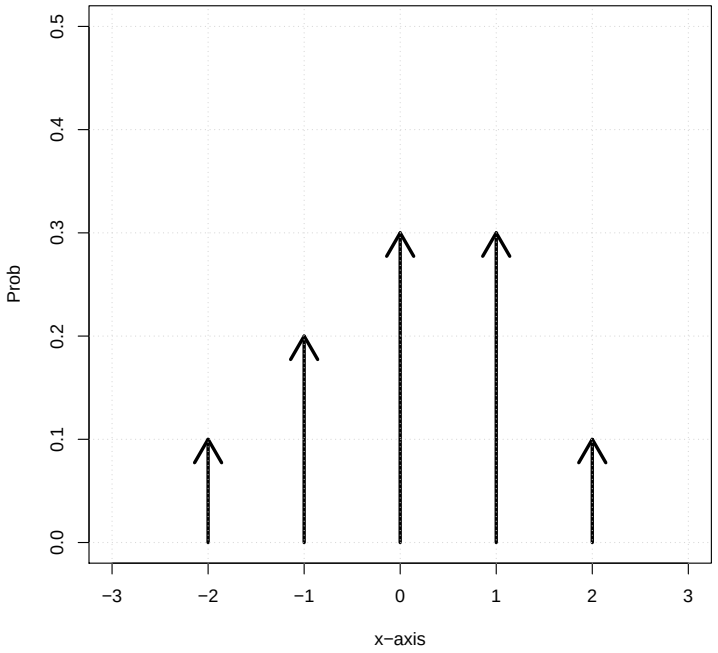

Figure 1.2: Discrete probability distribution for example 1

Example 1

Given the arbitrary discrete distribution shown in Fig. 1.2, calculate its moments and then reproduce the original distribution by the Von Mises method.

The first 10 moments of the distribution of Fig. 1.2 are presented in Table 1,

Table 1. Moments of the distribution presented in Fig. 1.2

Order	Moment
1	0.10000
2	1.30000
3	0.10000
4	3.70000
5	0.10000
6	13.3000
7	0.10000
8	51.70000
9	0.10000
10	205.3000

The first step in the Von Mises solution is to find whether or not those constants or a number of them, are indeed the moments of certain unknown distribution. To do this, the sign of a series of determinants should be examined, as presented below,

$$D_0 = 1 \tag{1.38}$$

$$D_1 = \begin{vmatrix} 1 & 0.1 \\ 0.1 & 1.3 \end{vmatrix} = 1.29 \tag{1.39}$$

$$D_2 = \begin{vmatrix} 1 & 0.1 & 1.3 \\ 0.1 & 1.3 & 0.1 \\ 1.3 & 0.1 & 3.7 \end{vmatrix} = 2.592 \tag{1.40}$$

$$D_3 = \begin{vmatrix} 1 & 0.1 & 1.3 & 0.1 \\ 0.1 & 1.3 & 0.1 & 3.7 \\ 1.3 & 0.1 & 3.7 & 0.1 \\ 0.1 & 3.7 & 0.1 & 13.3 \end{vmatrix} = 7.0848 \tag{1.41}$$

$$D_4 = \begin{vmatrix} 1 & 0.1 & 1.3 & 0.1 & 3.7 \\ 0.1 & 1.3 & 0.1 & 3.7 & 0.1 \\ 1.3 & 0.1 & 3.7 & 0.1 & 13.3 \\ 0.1 & 3.7 & 0.1 & 13.3 & 0.1 \\ 3.7 & 0.1 & 13.3 & 0.1 & 51.7 \end{vmatrix} = 14.9299 \tag{1.42}$$

$$D_5 = \begin{vmatrix} 1 & 0.1 & 1.3 & 0.1 & 3.7 & 0.1 \\ 0.1 & 1.3 & 0.1 & 3.7 & 0.1 & 13.3 \\ 1.3 & 0.1 & 3.7 & 0.1 & 13.3 & 0.1 \\ 0.1 & 3.7 & 0.1 & 13.3 & 0.1 & 51.7 \\ 3.7 & 0.1 & 13.3 & 0.1 & 51.7 & 0.1 \\ 0.1 & 13.3 & 0.1 & 51.7 & 0.1 & 205.3 \end{vmatrix} = -2.27813E-04 \tag{1.43}$$

the last postive determinant is D_4, hence a distribution can be determined with the first 8 moments of Table 1. It also means that the solution has 5 impulses (number of element in the determinant's columns). The next step is the solution of the system of linear eq. given by,

$$\begin{bmatrix} 1 & 0.1 & 1.3 & 0.1 & 3.7 \\ 0.1 & 1.3 & 0.1 & 3.7 & 0.1 \\ 1.3 & 0.1 & 3.7 & 0.1 & 13.3 \\ 0.1 & 3.7 & 0.1 & 13.3 & 0.1 \\ 3.7 & 0.1 & 13.3 & 0.1 & 51.7 \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \\ c_4 \end{bmatrix} = \begin{bmatrix} -0.1 \\ -13.3 \\ -0.1 \\ -51.7 \\ -0.1 \end{bmatrix} \tag{1.44}$$

The solution of this system gives the coefficients of the polynomial presented in eq. 1.32,

$$x^5 + 2.648E-06x^4 - 5x^3 - 1.14E-05x^2 + 4x + 4.336E-06 \tag{1.45}$$

The abscissae of the 5-step distribution are the roots of the polynomial above. They are presented in Table 2.

Table 2. X-values

Root	Value
1	-2.0 + 0.0i
2	2.0 + 0.0i
3	-1.0 + 0.0i
4	1.0 + 0.0i
5	-1.084E-06 + 0.0i

Finally the impulses corresponding to these abscissae are given by the solution of the Vandermonde matrix presented in eq. 1.33. The solution (p values) are presented in Table 3,

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ -2 & 2 & -1 & 1 & -1.1E-06 \\ 4 & 4 & 1 & 1 & 1.2E-12 \\ -8 & 8 & -1 & 1 & -1.3E-18 \\ 16 & 16 & 1 & 1 & 1.4E-24 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \\ p_5 \end{bmatrix} = \begin{bmatrix} 1.0 \\ 0.1 \\ 1.3 \\ 0.1 \\ 3.7 \end{bmatrix} \tag{1.46}$$

Table 3. Impulses of the distribution.

p_1	0.0999998
p_2	0.0999999
p_3	0.2
p_4	0.3
p_5	0.3000001

Note that the impulses correspond to the locations presented in Table 2, that is, p_1 corresponds to -2, p_2 corresponds to +2, etc. The final values of the discrete distribution calculated by the Von Mises method are presented in Table 4. Using normal numerical approximation i.e. $0.0999998 \approx 0.1 \; and \; -1.084E-06 \approx 0.0$ we can see that the method produces very accurate results.

Table 4. Calculated discrete distribution (ordered)

x-axis	Prob
-2	0.0999998
-1	0.2
-1.084E-06	0.3000001
+1	0.3
+2	0.0999999

Example 2

. Given the continuous distribution presented in Fig. 2 calculate the corresponding 3-point discrete distribution using the Hasofer method.

This distribution represents the maximum daily wind speed (in km/h) registered by the Australian Bureau of Meteorology at the Sydney Airport from 1980 to 2006 (BoM, 2017). The distribution has been calculated from the histogram of wind speed as the kernel density function. Table 5 presents some characteristics of the distribution.

Table 5. Characteristics of the distribution of Fig. 2

range	1980-2006
Max speed	114.9
min speed	0.0
number of observations	9671

Table 6. Moments of the distr. of Fig. 2

Moments		Central moments
1:	4.45E+01	0.0E+01
2:	2.23E+03	2.52E+02
3:	1.23E+05	1.75E+03
4:	7.42+06	2.05E+05

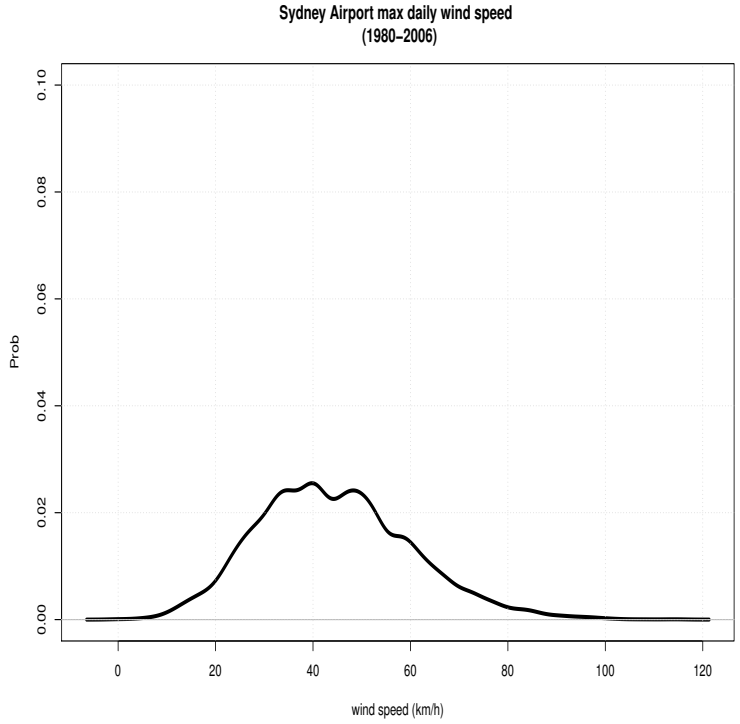

Figure 1.3: Prob density of wind speed at Sydney Airport

The first step in the solution of the problem is to calculate the moments of the distribution. Since we are working with a pdf the appropriate moments to calculate are the weighted moments given by the expression:

$$m_r = \sum_{\forall i} x_i^r \; p_i / \sum_{\forall i} p_i \tag{1.47}$$

Table 6 presents the first four moments of the distribution and its corresponding CM. Using these moments we can calculate the equivalent three-point distribution. Table 7 presents a summary of the calculations. The final discrete distribution is presented in Table 8.

Table 7. Summary of calculations

$\Delta =$	27.89727
x =	31.3656
y =	24.42894
p =	0.1441532
q =	0.1850858

Table 8. Three-point equivalent distribution

x-axis	Prob
20.0	0.185
44.0	0.671
75.8	0.144

In Fig. 3 the three-point equivalent distribution is plotted on top of the distribution of wind speed presented before.

The three-point equivalent distribution greatly simplifies problems involving random variables, see for example Sanabria et al. (1999, 2003).

1.6.2 Case B: We know that rv Z is a continuous distribution

.

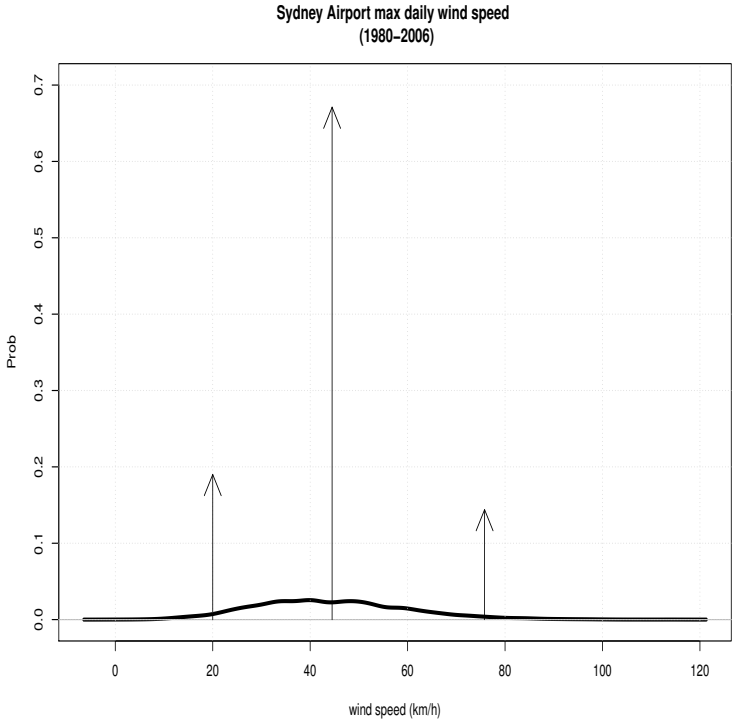

Figure 1.4: Three-point equivalent distribution

Fitting continuous distributions by the method of moments can be done using a variety of standard distributions. A well-known technique is by using the Pearson family of curves. This technique is presented in Chapter 6 of Kendall & Stuart (1950). In particular notice that a Pearson Type I distribution is the Beta distribution, Pearson Type III is the Gamma distribution, etc.

A more general technique, which is especially effective in practical applications, is fitting a Gram-Charlier distribution to the known moments. The idea behind this distribution is based on the properties of the Normal. If our set of parameters have only the first two moments, the other ones being 0 or close to 0, we know that we can fit a Normal distribution with these two moments. If moments higher that 2 are not 0 then the way the unknown distribution differs from the Normal is a proportion of these parameters. For this reason the distribution is a series expansion around the Normal.

Recall the standard Normal distribution,

$$N(z) = (1/\sqrt{(2\pi)}) \exp(-z^2/2) \tag{1.48}$$

where z is the standardized x-axis point,

$$\begin{gathered} \text{z} = (\text{x - } m_1)/\sigma \\ m_1 = \text{Mean of the distribution} \\ \sigma = \text{standard deviation} \end{gathered}$$

The Gram-Charlier series is given by,

$$f(x) = N(x) + \sum_{i=3}^{\infty} \frac{\kappa_i}{i!} N^i(x) \tag{1.49}$$

where $N^i(x)$ is the i-th derivative of the Normal distribution. The derivative of the Normal can be expressed through the Hermite polynomials 'He' as,

$$N^i(x) = (-1)^i He_i N(x) \tag{1.50}$$

Hence the Gram-Charlier series becomes,

$$f(x) = N(x)\{1 + \kappa_3 H_3/3! + \kappa_4 H_4/4! + \kappa_5 H_5/5! + (\kappa_6 + 10\kappa_3^2)/6! + ...\} \tag{1.51}$$

where H_r is the Hermite polynomial of order 'r' and κ_r is the cumulant of order 'r'. The expansion of the expression above have given rise to different distributions termed Gram-Charlier series A and B; and Edgeworth series (Cramer, 1957). In practical work only 3 or 4 terms are used.
The Hermite polynomial of order 'r+1' can be found using,

$$H_{r+1}(z) = zH_r(z) - rH_{r-1}(z) \tag{1.52}$$

starting with,

$$H_0 = 1$$
$$H_1 = \text{z}$$
$$H_2 = z^2 \text{ - } 1$$

The corresponding cumulative distribution is given by,

$$F(x) = \int_{-\infty}^{x} f(s)ds = \int_{-\infty}^{x} N(s)ds + \sum_{i=3}^{\infty} \frac{\kappa_i}{i!} N^{(i-1)}(x) \tag{1.53}$$

Example 3

Fit a Gram-Charlier series to the distribution of wind speed presented in Fig. 1.3. The first step in the solution of the problem is to calculate the moments and cumulants of the wind distribution. These parameters can be calculated using eq (4) and (7) and are presented in Table 9,

Table 9. Moments and Cumulants of the distr. of Fig. 2

	Moments	Cumulants
1:	4.45E+01	4.45E+01
2:	2.23E+03	2.52E+02
3:	1.23E+05	1.75E+03
4:	7.42E+06	1.45E+04
5:	4.80E+08	1.11E+04
6:	3.32E+10	8.06E+06

Fig. 1.5 shows the prob density of wind speed using the Gram-Charlier series (GC-series)(in red). Note that the series extends below positive values, for this reason the prob density of this funcion is seldom used (Straja 2020), the distribution function will be extensively used in Section 2.2.1.

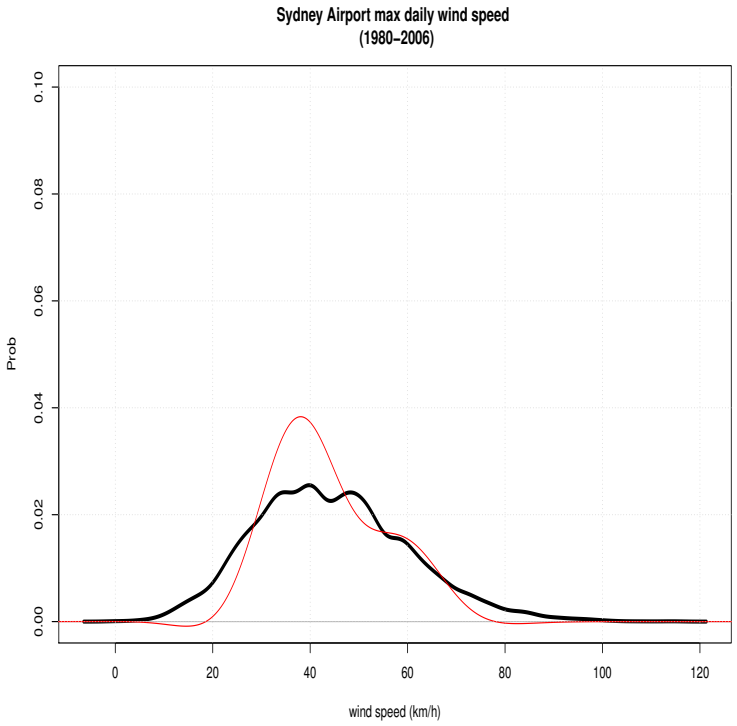

Figure 1.5: Kernel density and GCh-series of wind speed

1.6.3 Case C: We know that rv Z is a mix of continuous and discrete distributions

.

This case is more complicated. If the input variables to the model comprise continuous and discrete distributions it is necessary to solve the problem twice: one for the continuous distributions and another for the discrete distributions. In this case the solution of the problem using eq. 1.9 gives two sets of cumulants κ_r, one set comes from the continous distributions and the other comes from the discrete distributions, in other words we have r cumulants κ_{r_c} and r cumulants κ_{r_d}. The pdf of rv Z would be a combination of discrete and continuous distributions.

To simplify things let's assume that the continuous part of the problem is modelled only with Normal distributions. Because of the reproductive property of the Normal we know that the cumulants κ_{r_c} produce a single normal distribution with parameters m_{1_c} and σ_c, these parameters are calculated from,

$$\begin{aligned} m_{1_c} &= \kappa_{1_c} \\ \sigma_c &= \sqrt{\kappa_{2_c}} \end{aligned} \tag{1.54}$$

Similarly the cumulants of the discrete distribution produce a single discrete distribution, this discrete distribution is calculated from the cumulants of the discrete part of the solution κ_{r_d} using the VMSF as explained in Section 1.6.1. The convolution of the two distributions is illustrated in Fig. 1.6 (Sanabria and Dillon, 1986). Mathematically,

$$\mathrm{f(z)} = (1/\sigma_c\sqrt{2\pi})\{p_1 \exp\left(-\tfrac{[z-(y_1+m_{1_c})]^2}{2\sigma_c^2}\right) + p_2 \exp\left(-\tfrac{[z-(y_2+m_{1_c})]^2}{2\sigma_c^2}\right) + \ldots p_n \exp\left(-\tfrac{[z-(y_n+m_{1_c})]^2}{2\sigma_c^2}\right)\} \tag{1.55}$$

1.7 Correlation

Correlation between random variables can be easily included in the model by modifying the calculation of the Variance (second cumulant) in eq. 1.12. Let us suppose that rvs X and Y are not statistically independent hence the second cumulant of eq. 1.12 becomes,

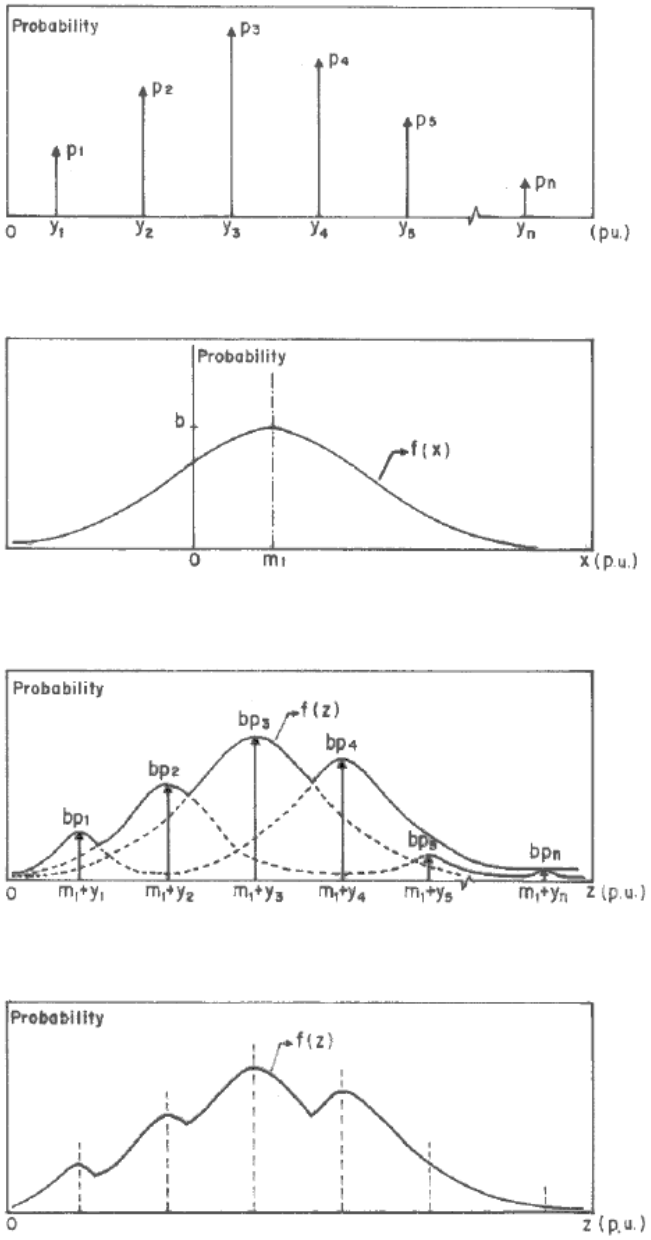

Figure 1.6: Convolution of a discrete rv and a Normally distributed rv

$$\kappa_2 = a^2\kappa_2(x) + b^2\kappa_2(y) + 2\ cov(x,y) \tag{1.56}$$

where cov(x,y) is the covariance between random variables X and Y.

The covariance between rv X and rv Y indicates whether both variables follow the same pattern of change. If X tends to increase or decrease along with Y, the covariance would be large and positive. If, on the other hand, Y tends to increase as X decreases their covariance would be a large negative number.

In practice the covariance of rvs X and Y is calculated from the definition of correlation coefficient ρ_{xy},

$$\rho_{xy} = cov(x,y)/(\sigma_x\sigma_y) \tag{1.57}$$

that is,

$$cov(x,y) = \rho_{xy} * (\sigma_x\sigma_y) \tag{1.58}$$

The correlation coefficient lies between -1 and +1. Values close to 1 show a linear correlation between rv X and rv Y.

Eq. (1.50) above can be extended to the case of "n" non-independent rvs. Consider the expression,

$$Z = X_1 + X_2 + X_3 + \ldots X_n \tag{1.59}$$

The Variance of Z (second Cumulant) becomes,

$$\mathrm{V\{Z\}} = \mathrm{V\{X_1\}} + \mathbf{V}\{X_2\} + \mathbf{V}\{X_3\} + \ldots \mathbf{V}\{X_n\} + 2\ \textstyle\sum_{\forall i}\ \sum_{\forall j}\ cov(x_i\ x_j\ )$$

where $\forall i$ means for i = 1 to n-1; and $\forall j$ means for j = i+1 to n

1.8 References

Kendall & Stuart (1950). The Advanced Theory of Statistics. Vol I. Charles Griffin & Co. Ltd. London.

Schellenbarg A., Rosehart W. and Aguado J. (2005). Cumulant-based Probabilistic Optimal Power Flow with Gaussian and Gamma Distributions. IEEE Trans. on Power Systems. Vol. 20, No. 2, May 2005.

Breitenberger E. (1991). Probability, convolutions, and distributions: EPRI monographs on simulation of electric power production. Report No. EPRI-IE–7508. Research Reports Center, PO Box 50490, Palo Alto, CA 94303.

Sanabria L.A., Soh B., Dillon T.S., Chan E. (2003). Reliability Optimisation of a Web Server. Proceedings of the International Conference on Internet and Multimedia Systems and Applications (IMSA 2003). Hawaii. August 13-15, 2003.

Sanabria L.A. (1999). Human Behaviour in Fires -Computer Program and Results. RISK'99 'Back to the Future', Risk Engineering Society Conference. Melbourne, August 1999.

Sanabria L.A., Dillon T.S. (1986). Stochastic Power Flow using Cumulants and Von Mises functions. Electrical Power Energy Systems. Vol. 8, No. 1, January 1986.

Bureau of Meteorology (BoM). Weather & climate data. Climate data online (accessed on 20/02/2017).

Lukacs E. (1960). Characteristic Functions. Griffins Statistical Monographs Courses, Charles Griffin Co. London.

Thiele T.N. (1931). The Theory of Observations.Annals of Mathematical Statistics, Vol. 2, pp165.

Von Mises R. (1964). Mathematical Theory of Probability and Statistics. Academic Press, New York.

Hasofer A.M. (1997). Three-point representation of a distribution. CESARE Report, VUT. June 1997.

Straja S.R. Application of the Gram-Charlier Approximation for Option Valuation. In: http://www.fintools.com/wp-content/uploads/2012/02/DerivativeValuation-GramCharlier.pdf accessed on 26/04/2020.

Cramr H. (1957). Mathematical Methods of Statistics. Princeton University Press, Princeton.

Chapter 2

Realistic sized example problems

To illustrate the technique presented in the previous chapter, a couple of realistic sized example problems will be solved in this chapter. The first problem is a fundamental problem in electrical power systems: the flow of power in an electric power network. This problem is more commonly known as the 'load flow' problem. The second problem is known as the 'probabilistic production costing' problem. In this problem we want to know the cost of producing electrical energy in a power system to meet the demand of electricity. The generating plants are subjected to random outages and the demand varies at random from season to season and during the day. Hence it is important to model the problem with random variables. Let us start with the first problem.

2.1 Probabilistic Load Flow

The power system used in the example is the IEEE 14-busbar, 20-line system. The power system layout is presented in Fig. 1.7. The network data is presented in Table 10.

Table 10. Network data for the IEEE 14-busbar, 20-line power system

Busbar Sending	Busbar Receiving	Resistance p.u.	Reactance p.u.	Susceptance p.u.	Transformer Tap (%
1	2	0.01938	0.05917	0.0264	-
1	5	0.050403	0.22304	0.0264	-
2	3	0.04699	0.19797	0.0219	-
2	4	0.05811	0.176632	0.0187	-
2	5	0.05695	0.17388	0.0170	-
3	4	0.06701	0.17103	0.0064	-
4	5	0.01335	0.04211	-	-
4	7	-	0.20912	-	-2.2
4	9	-	0.55618	-	-3.1
5	6	-	0.25202	-	-6.8
6	11	0.09498	0.19890	-	-
6	12	0.12291	0.25581	-	-
6	13	0.06615	0.13027	-	-
7	8	-	0.17615	-	-
7	9	-	0.11001	-	-
9	10	0.03181	0.08450	-	-
9	14	0.12711	0.27038	-	-
10	11	0.08205	0.19207	-	-
12	13	0.22092	0.19988	-	-
13	14	0.17093	0.34802	-	-
9	9	-	-5.260	-	-

The input data for the conventional or 'deterministic' solution of the problem consists of the injectec active powers at all nodes except the 'slack' node, the injected reactive powers at all load or PQ nodes and the voltage at all generator or PV nodes. The nodal input data are related to the nodal angles and voltages by the non-linear eq. 2.1,

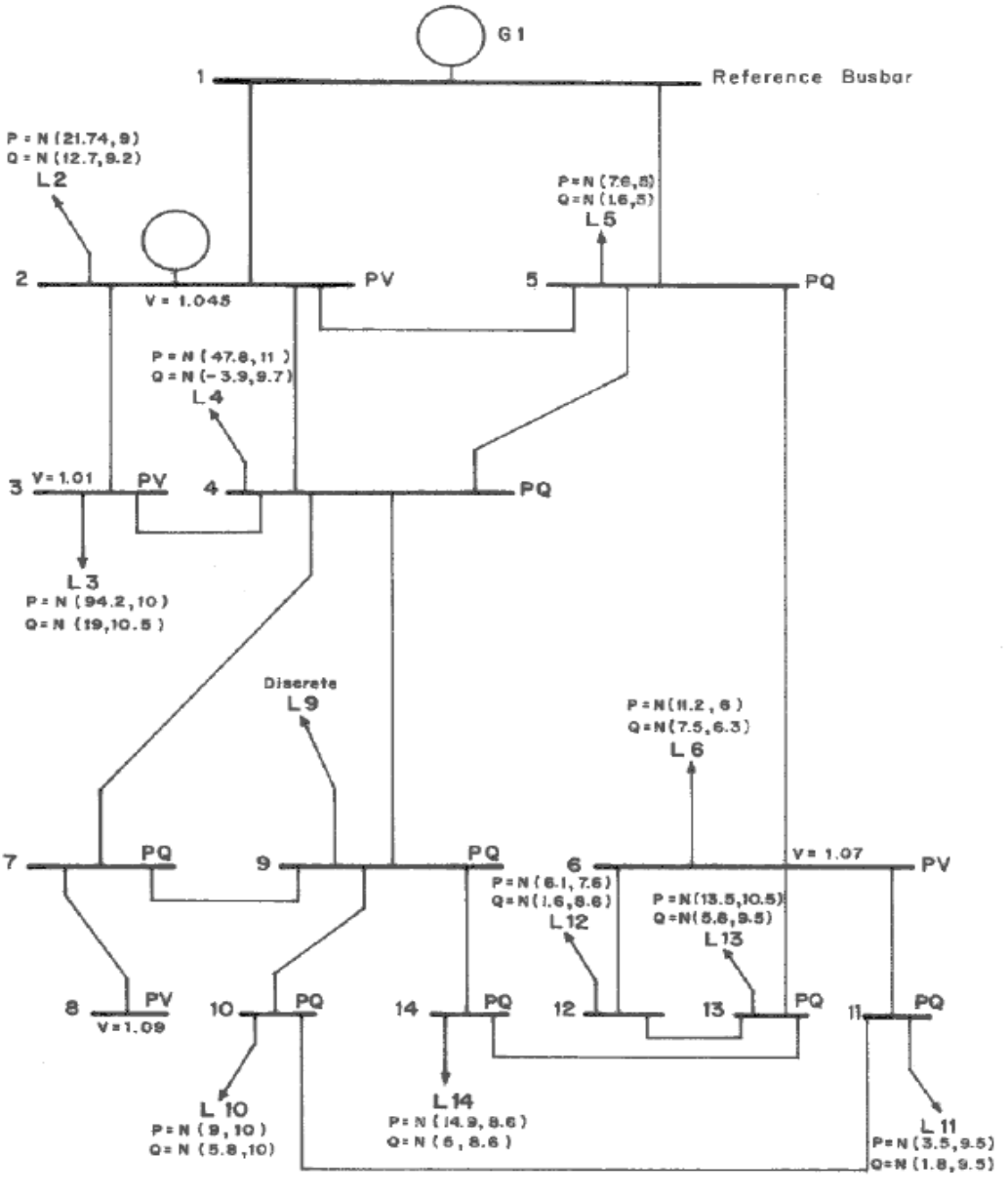

Figure 2.1: Layout of the IEEE 14-busbar, 20-line network

$$
\begin{aligned}
P_i &= V_i \sum_{j=1}^{n} V_k(G_{ik}cos\theta_{ik} + B_{ik}sin\theta_{ik}) \\
Q_i &= V_i \sum_{j=1}^{n} V_k(G_{ik}sin\theta_{ik} - B_{ik}cos\theta_{ik}) \\
for \quad & i = 1, 2, 3 \cdots n \quad (nodes)
\end{aligned}
\tag{2.1}
$$

With the angles and voltages of the nodes known, it is possible to calculate the load flow in the branches using the power-flow equations,

$$
\begin{aligned}
P_{ik} &= -t_{ik}G_{ik}V_1^2 + V_iV_k(G_{ik}cos\theta_{ik} + B_{ik}sin\theta_{ik}) \\
Q_{ik} &= t_{ik}B_{ik}V_1^2 - B'_{ik}V_i^2 + V_iV_k(G_{ik}sin\theta_{ik} + B_{ik}cos\theta_{ik})
\end{aligned}
\tag{2.2}
$$

where,

$$
\begin{aligned}
t_{ik} &= \text{transformer tap ratio,} \\
V_i &= \text{voltage at node i,} \\
\theta_{ik} &= \text{difference in voltage angle between nodes i and k,} \\
G_{ik} &= \text{real part of element ik of admittance matrix,} \\
B_{ik} &= \text{imaginary part of element ik,} \\
B'_{ik} &= \text{half of susceptance of line ik}
\end{aligned}
$$

The input data for the probabilistic solution of the problem is the same than for the deterministic one. However, in the former the variables are not single-value quantities but random variables. Modelling the system with rv introduces difficulties which can be summarised as follows:

- The random variables P,Q, θ and V are not necessarily independent from each other.
- The variables θ and V are also rv because they depend on P_i and Q_i. Additionally. some mathematical manipulation must be done to relate explicity θ and V with P and Q.
- The problem is further complicated by the fact that eq. X and Y are non-linear.

Because of these complications a number of assumptions have to be made in order to solve the problem:

- The random variables P and Q are considered independent from each other and hence θ and V are also independent.
- The solution considers a 'decoupled' power system, i.e. V depends only on P while the voltage angle depends on Q.
- Finally, the power-flow eq. needs to be linearised to be able to solve the problem using eq. 1.12. A simple technique to linearise the power-flow eq. is presented below.

2.1.1 Linearisation of the power-flow equations

A linearisation technique particularly appropriate for power systems is due to Allan and Al-Shakarchi (1977). In this technique the linearisation is developed considering that the random nature of a variable is due to random changes around the deterministic value. The rv X can, then, be written as,

$$X = X_o + \Delta X \tag{2.3}$$

where, X_o = Mean (Expected value) of the rv. This value should coincide with the value of X in the deterministic problem. ΔX = Random changes around X_o

Similarly the rv Y is given by,

$$Y = Y_o + \Delta Y \tag{2.4}$$

The multiplication of rvs X and Y is then given by the expression,

$$Z = X.Y = X_oY_o + X_o\Delta Y + Y_o\Delta X + \Delta X\Delta Y \tag{2.5}$$

If the random variations around the mean are small, the last term can be neglected, i.e.

$$Z = X.Y = X_o Y_o + X_o \Delta Y + Y_o \Delta X \tag{2.6}$$

Using this technique the injection of active power at node i (eq. 2.1 above) can be expressed as,

$$P_i = V_{i_o} \sum_{j=1}^{n} V_{k_o} (f_{ik}\theta_i - f_{ik}\theta_k + e_{ik}) \tag{2.7}$$

of course we are interested in the unknown, the angle of the voltage at node i, so by inverting P_i we obtain,

$$\theta_i = \sum_{j=i}^{n-1} \hat{Y}_{ij} P_j - \sum_{j=i}^{n-1} \hat{Y}_{ij} R_j \tag{2.8}$$

where,

$\hat{Y}_{ij}$ = (i,j)-th element of the inverted admittance matrix
P_j = rv of net injected active power at node j
θ_i = rv of voltage angle at node i
R_j = constant

Note that eq. (1.61) has the same form of eq 1.1, that is,

$$\boldsymbol{\theta} = \boldsymbol{Y}\boldsymbol{P} + \boldsymbol{C} \tag{2.9}$$

where the bold font indicates matrix, $\boldsymbol{Y}$ and $\boldsymbol{C}$ are constants, $\boldsymbol{P}$ is the rv.

In the decoupled model, the flow of active power in line i-k depends only on the angles of the node voltage, i.e.

$$P_{ik} = g_{ik}\theta_i - g_{ik}\theta_k + h_{ik} \tag{2.10}$$

where.

θ_i and θ_k are rv
P_{ik} = rv of active power flow on line connecting nodes i-k
h_{ik} = constant

Similarly for the flow of reactive power: linearisation of eq. 2.1 gives,

$$Q_i = \sum_{k=1}^{n} A_{ik}(V_{k_o} V_i + V_{i_o} V_k - V_{i_o} V_{k_o}) \tag{2.11}$$

By inversion the voltages in the PQ-type nodes and the net injected reactive power in the PV-type busbars can be found. From these voltages the flow of reactive power between nodes i-k can be calculated by the expression,

$$Q_{ik} = \alpha_{ik} V_i + \beta_{ik} V_k + \gamma_{ik} \tag{2.12}$$

A detailed explanation of the variables in these equations is presented in Allan and Al-Shakarchi (1979) and Sanabria and Dillon (1986).

For the probabilistic load-flow (PLF) model let the input data consists of two generating plants connected to nodes 1 and 2. The plant in node 1 has 11 identical units, each one with a Rating Capacity of 25 MW and Forced Outage Rate (FOR) of 8%. The plant of node 2 has 3 identical units of 22 MW each and FOR of 9%. Both plants can be modelled with Binomial distributions. The loads of the system are modelled using Normal distributions except for the load of node 9 which is represented by two discrete functions of 5 impulses each, as shown in Fig. 2.2.

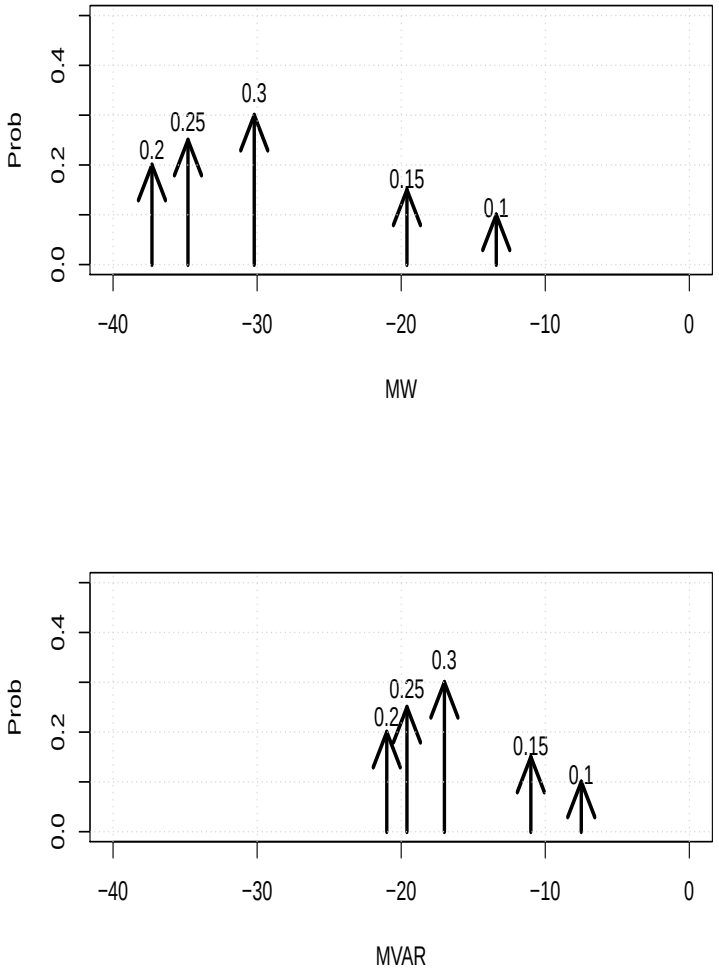

Figure 2.2: Discrete rv of active and reactive power in node 9

Table 11 shows the cumulants of the net input data for solution of the active part of the PLF problem. 'N' in the table refers to the Normal, 'Bin' refers to the Binomial distribution. Note that nodes modelled with Normal distributions have only 2 cumulants κ_1, κ_2, the other cumulants are 0.0. For the discrete part we are calculating 9 cumulants but to save space only 4 are presented in Table 10. All units are 'per unit' or p.u. (Base = 100 MW).

Table 11. Cumulants of input data (active part)

Node	type	Model	κ_1	κ_2	κ_3	κ_4
1	Gen	Bin	2.3	4.6E-02	-9.7E-03	1.6E-03
2	Gen	Bin	0.4	8.0E-03	-1.4E-03	2.0E-4
2	Load	N	2.2E-01	3.8E-04	0.0	0.0
3	Load	N	9.4E-01	8.9E-03	0.0	0.0
4	Load	N	4.8E-01	2.8E-03	0.0	0.0
5	Load	N	-7.6E-02	1.4E-05	0.0	0.0
6	Load	N	1.1E-01	4.5E-05	0.0	0.0
10	Load	N	9.0E-02	8.1E-05	0.0	0.0
11	Load	N	3.5E-02	1.1E-05	0.0	0.0
12	Load	N	6.1E-02	2.2E-05	0.0	0.0
13	Load	N	1.4E-01	2.0E-04	0.0	0.0
14	Load	N	1.5E-01	1.7E-04	0.0	0.0

2.1.2 Solution of the power-flow equations

As explained in Section 1.6.3, eq. (1.61) has to be solved twice: one using only the cumulants of the discrete part of the problem and the other one using the cumulants of the continuous part. In other words, the first solution of the problem uses only the data from nodes 1 and 2 of Table 11, the second solution uses the data from nodes 2 to 14 to produce the continuous-type cumulants. Solution of eq (2.8) produces the cumulants of θ, the angle of the voltages, as presented in Table 12.

Table 12. Cumulants of voltage angle in radians
(continuous cumulants are given first)

Node	cumulant type	κ_1	κ_2	κ_3	κ_4
2	cont.	-9.8E-02	2.8E-05	0.0	0.0
2	discr.	1.7E-02	3.1E-05	-1.6E-07	1.3E-09
3	cont.	-2.3E-01	2.5E-04	0.0	0.0
3	discr.	1.9E-02	4.6e-05	1.1e-08	4.4e-10
4	cont.	-1.7E-01	8.4E-05	0.0	0.0
4	discr.	6.8e-03	6.8e-05	2.9e-07	-9.3e-10
5	cont.	-1.5E-01	5.5E-05	0.0	0.0
5	discr.	8.1e-03	5.2e-05	1.7e-07	-4.3e-10
6	cont.	-2.2E-01	8.1E-05	0.0	0.0
6	discr.	-1.1e-02	1.3e-04	1.2e-06	-6.7e-09
7	cont.	-2.0E-01	8.6E-05	0.0	0.0
7	discr.	-1.8e-02	2.2e-04	2.6e-06	-1.9e-08
8	cont.	-2.0E-01	8.6E-05	0.0	0.0
8	discr.	-1.8e-02	2.2e-04	2.6e-06	-1.9-08
9	cont.	-2.2E-01	8.9E-05	0.0	0.0
9	discr.	-3.1E-02	3.3E-04	5.1E-06	-4.6E-08
10	cont.	-2.3E-01	0.1E-01	0.0	0.0
10	discr.	-2.5e-02	2.9e-04	4.2e-06	-3.5e-08
11	cont.	-2.3E-01	8.6E-05	0.0	0.0
11	discr.	-1.6e-02	2.1e-04	2.4e-06	-1.7e-08
12	cont.	-2.4E-01	8.9E-05	0.0	0.0
12	discr.	-6.3e-03	1.4e-04	1.3e-06	-7.8e-09
13	cont.	-2.5E-01	9.7E-05	0.0	0.0
13	discr.	-4.7e-03	1.6e-04	1.5e-06	-9.5e-09
14	cont.	-2.6E-01	1.1E-04	0.0	0.0
14	discr.	-1.1e-02	2.5e-04	3.2e-06	-2.5e-08

Once the cumulants of the voltage angles have been calculated we can calculate the cumulants of

the flow of active power in the lines of the power system using eq. (2.10). Table 13 presents these cumulants for the first 10 lines.

Table 13. Cumulants of active power flow in p.u.
(continuous cumulants are given first)

Line	between nodes	cumulant type	κ_1	κ_2	κ_3	κ_4
1	1-2	cont.	1.7	8.5E-03	0.0	0.0
1	1-2	discr.	-0.24	9.3E-03	8.2E-04	1.1E-04
2	1-5	cont.	0.7	1.2E-03	0.0	0.0
2	1-5	discr.	-7.2e-04	1.1e-03	-1.8e-05	-2.1e-07
3	2-3	cont.	0.7	3.1E-03	0.0	0.0
3	2-3	discr.	2.5e-02	1.3e-04	-1.3e-06	-7.3e-09
4	2-4	cont.	0.4	5.4E-04	0.0	0.0
4	2-4	discr.	1.0e-01	5.9e-04	-1.2e-05	-1.4e-07
5	2-5	cont.	0.3	2.0E-04	0.0	0.0
5	2-5	discr.	9.5e-02	3.7e-04	-6.1e-06	-4.1e-08
6	3-4	cont.	-2.9E-01	1.9E-03	0.0	0.0
6	3-4	discr.	4.6e-02	1.2e-04	-1.1e-06	-5.6e-09
7	4-5	cont.	-6.1E-01	1.6E-03	0.0	0.0
7	4-5	discr.	-4.3E-02	5.8E-04	1.1E-05	-1.3E-07
8	4-7	cont.	1.5E-01	5.4E-05	0.0	0.0
8	4-7	discr.	1.3E-01	1.3E-03	-4.2E-05	-7.6E-07
9	4-9	cont.	8.4E-02	1.8E-05	0.0	0.0
9	4-9	discr.	7.5E-02	4.2E-04	-7.9E-06	-8.2E-08
10	5-6	cont.	3.6E-01	1.6E-04	0.0	0.0
10	5-6	discr.	8.7E-02	4.6E-04	-9.3E-06	-1.0E-07

From the cumulants of the power flow it is possible to calculate the actual distribution function using the methodology explained in Section 1.6.3: first calculate the discrete function using the cumulants of the discrete part via the Von Mises Step Function. Then, from the continuous cumulants calculate the continuous function as a Normal with parameters m_{1_c} and σ_c, see eq. (1.49). The convolution of both distributions is given by eq. (1.50).

To illustrate the methodology consider the flow of active power in line 10, between nodes 5 and 6 (See Table 13). The cumulants of the discrete part (9 cumulants were calculated) produce the VMST shown in Table 14 and eq. (1.50) produces the density functions shown in Fig. (2.3): the discrete distibution is at the top, next is the comtinuous distribution; the convolution of these two distributions is shown at the bottom. The final result is the envelope of the last distribution as shown in Fig. (2.4). Note that the y-axis of this plot has been scaled by dividing by the area under the curve, since the area under the curve must add up to 1.0.

Table 14. Discrete distribution of active power flow in line 10
(between nodes 5 and 6)

i	x_i	$p(x_i)$
1	0.107	0.378
2	0.090	0.383
3	0.056	0.172
4	0.040	0.067

2.2 Probabilistic Production Costing

The Probabilistic Production Costing technique simulates the operation of a power system in order to meet the load demand. The usual simulation time is a week, but longer periods like seasonal or anual operation can also be simulated.

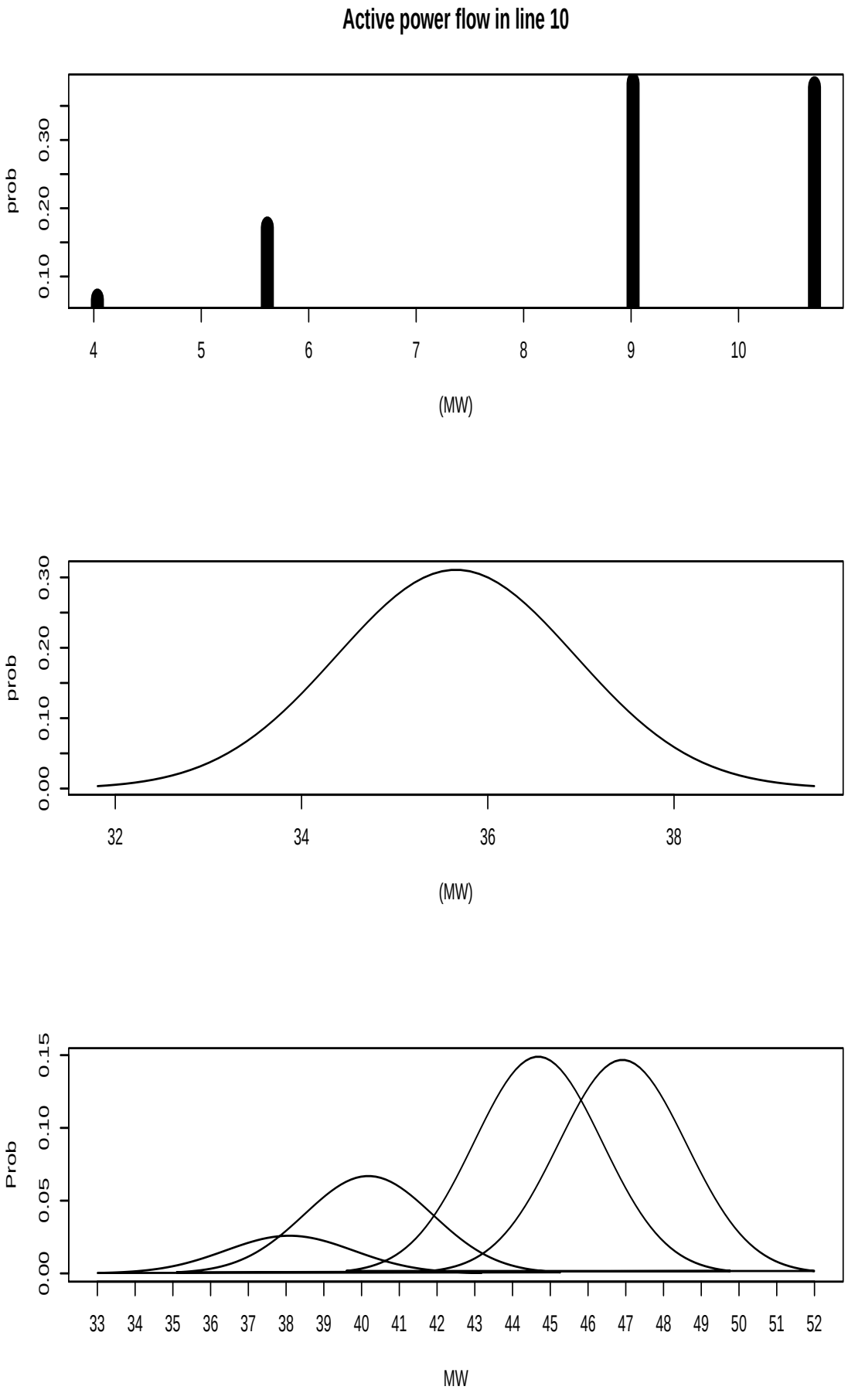

Figure 2.3: Calculation of the active power flow in line 10

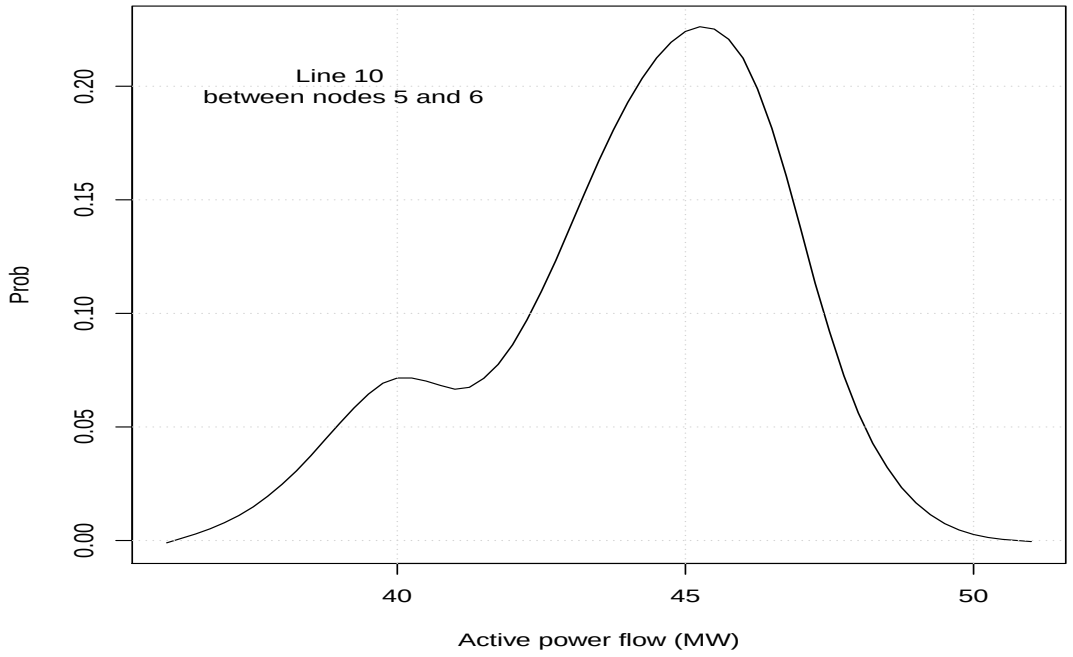

Figure 2.4: Active power flow in line 10 between nodes 5 and 6

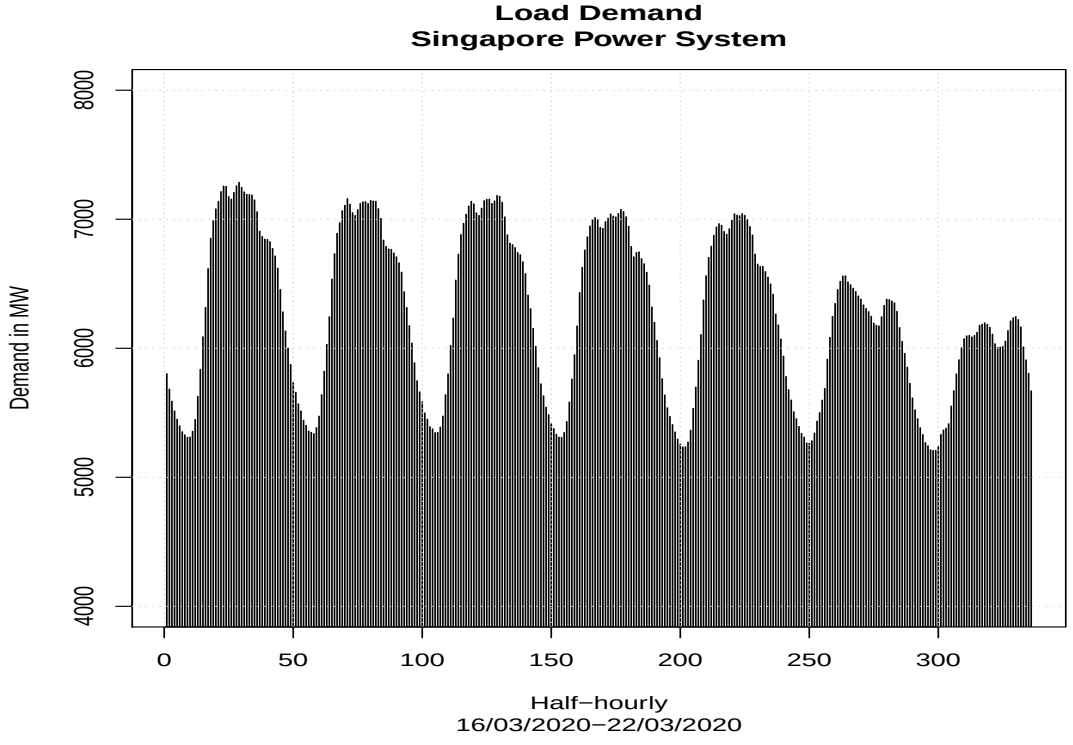

Figure 2.5: Weekly Load Demand of the Singapore power system

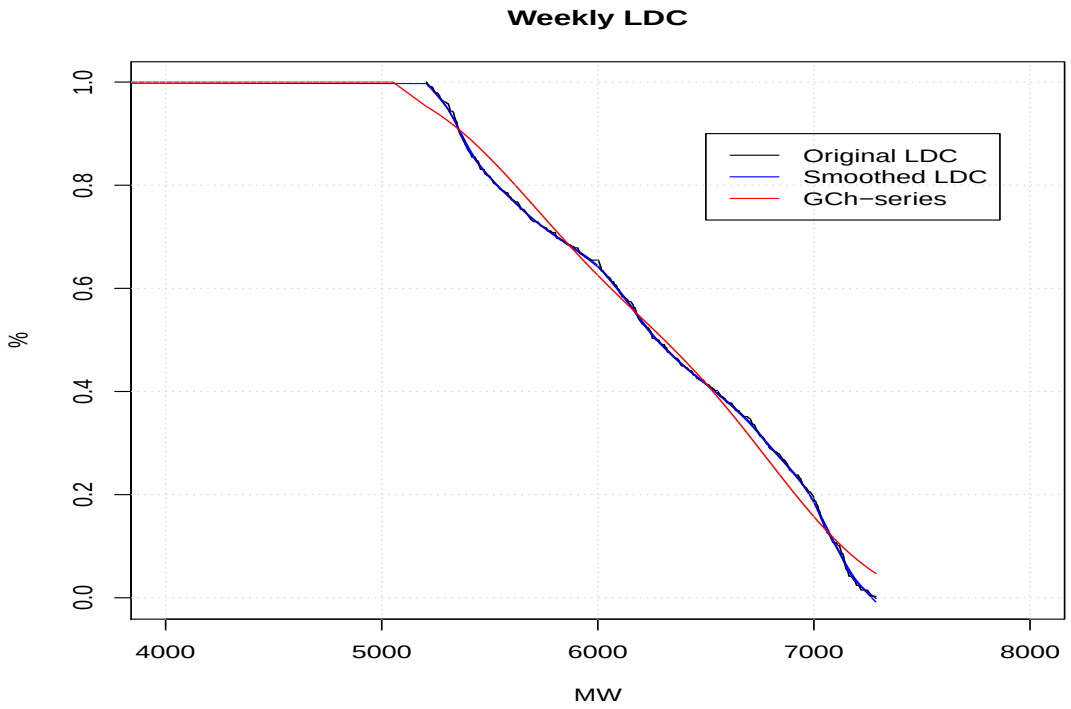

Figure 2.6: Corresponding Weekly Load Duration Curve

2.2.1 Probabilistic model for demand

The probabilistic model for the load demand can be developed by considering the chronological time series. The demand varies at random during the day and also during the week so it is necessary to treat this demand as a random variable. For this the demand can be transformed into an Inverted Load Duration Curve, this probabilistic curve gives us the probability that a certain value of the demand is exceeded (Finger 1979; Contaxi et al. 2003). Fig. 2.5 shows the weekly demand in the power system of Singapore in periods of half-hour, the minimum and maximum values are 5208 and 7286 (MW) respectively. The inverted LDC can easily been constructed using the Exceedance distribution, introduced in Section 1.4. this distribution is defined as,

$$Exc_{Sin} = 1.0 - CDF_{Sin} \tag{2.13}$$

where CDF_{Sin} is the empirical Cumulative distribution function of the Demand of Fig. 2.5. The inverted LDC is, by definition, the Exc_{Sin}, this distribution is presented in Fig. 2.6. Note that the probability of exceeding values of demand between 0.0 and 5208 (the minimum demand) is 1.0, i.e. values between 0.0 and minimum demand are always exceeded. On the other hand the probability of exceeding values of 7286 MW is 0.0, this value is never exceeded. In the cumulants method the inverted LDC is calculated from the cumulants of the load demand using the Gram-Charlier series expansion as explained in Section 1.6.2. The red curve in Fig. 2.6 shows the GCh-series. As mentioned in Section 1.6.2 the prob density function of the GCh-series does not have good fitting characteristics because of its tendency to extend beyond positive values, the LDC however has excellent fitting characteristics as shown in Fig. 2.6.

The importance of the inverted LDC is that the area under the curve beyond point "G1" is the Unserved Energy of the power system after generating plant "G1" has been loaded.

The mathematical model used to represent the generation plants are discrete distributions. To simplify the problem these are often represented by a two-state distribution. The increasing tendency to build large capacity generators, with the attendant risk of failure of boilers and auxiliary equipment, allows the derated operation of these units. The two-state model is, therefore inadequate for representation of fossil fired and nuclear plants and a multi-state and multi-block representation is often necessary. The multi-block representation comes from the fact that in operation it is seldom economical to commit one unit fully before another unit is loaded, rather the plants are committed in segments of similar marginal cost, and these segments are loaded, normally at non-adjacent positions, according to an economic order.

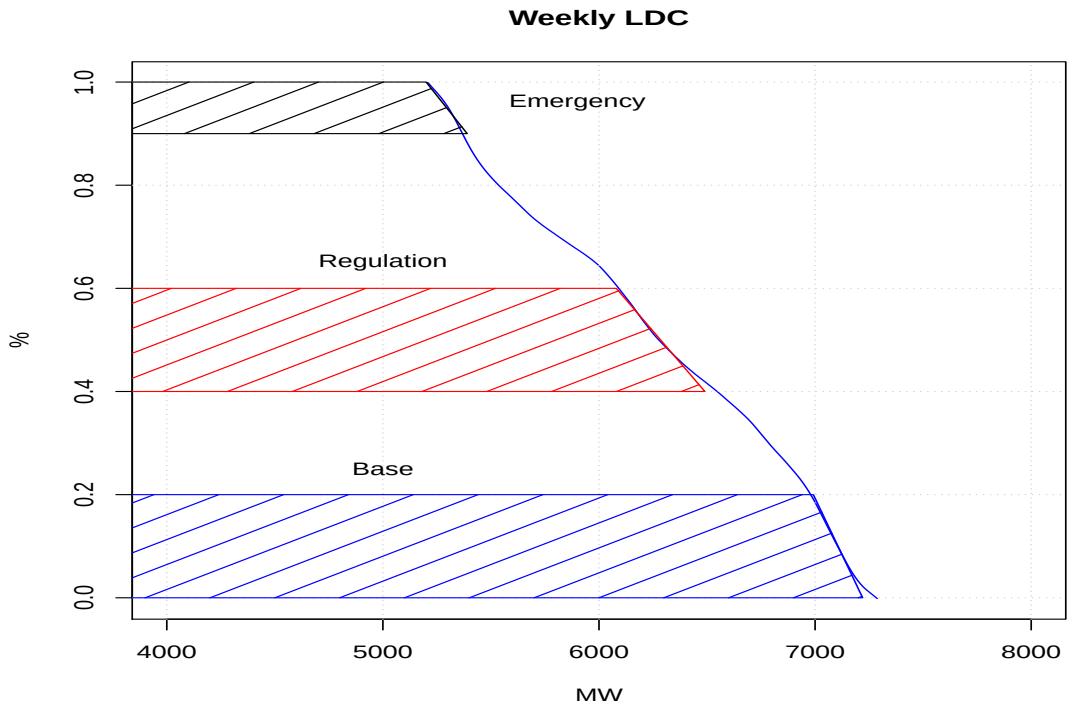

Figure 2.7: Hydro blocks loaded under the LDC

In systems with large amounts of energy generated by hydro plants the multi-block representation is specially important. In these systems any single hydro unit can be split into three or more blocks and individual blocks can be lumped together. The hydro energy basically consists of three blocks: Run of river or base load, normally associated with the firm capacity of the plant, regulating hydro and emergency hydro. Fig. 2.7 shows the position of the hydro blocks under the LDC for a power system with large hydro generation. It is also possible to separate the regulating blocks into several different regulating hydro sub-blocks.

2.2.2 Probabilistic model for generating plants

As mentioned before the generating plants can be modelled using discrete distributions. In this section we will consider only general discrete distributions. Extension to standard distributions such as binomial, normal etc. is trivial. Fig. 2.8 shows a general representation of a large hydro plant with the three blocks 'Hb', 'Hr' and 'He' clearly specified. We can see that the plant has a probability 'P4' of generating 'C4' MW.

The problem in this case is to find the best position of the regulating block to make the most efficient use of the available hydro energy. To fit this energy under the curve, some thermals, not necessarily complete units, can be de-loaded. Once the position of this hydro energy is found its equivalent capacity is calculated. Finally, the expected energies of the units or blocks of units on line and their costs are calculated.

In the probabilistic simulation the equivalent demand on a unit is defined as the sum of the customer demand and the demand due to failure of the plants lower in the merit order, i.e. due to plant failure, the next generation unit to be loaded should see a higher LDC, this is the so-called Equivalent Load Duration Curve (ELDC). Mathematically the ELDC is the convolution of demand and the Forced Outage Rate (FOR) of the units already dispatched (Finger 1979).

From Fig. 2.8 the corresponding FOR can be calculated and consequently the moments and the cumulants of each block can be found. From eq. (1.4) the moments can be calculated as,

$$m_k = \sum_{i=1}^{n} (C_i)^k q(C_i) \tag{2.14}$$

where,
$p(C_i)$ = Prob. of having a generation of C_i MW
$q(C_i)$ = Forced Outage Rate = 1.0 - $p(C_i)$
n = number of FOR states for the plant

2.2.3 Convolution

The moments can be transformed to cumulants using eq (1.8). One of the advantages of the method of cumulants is that the convolution of generating units' FOR and load demand can be done directly using the time series of demand withoug having to calculate the ELDC. Recall from eq. (1.1) and (1.10) that the convolution of two random variables is simply the sum of their cumulants.

As explained in Section 1.6.2, from the cumulants it is possible to calculate the corresponding probability distribution function 'F(zl)'. This function gives the probability of having a load of up to 'L' MW or 'zl' after normalizing, that is,

$$F(zl) = Pr[Z \leq zl] = \int_{-\infty}^{zl} f(s)ds \tag{2.15}$$

We are more interested in the complementary event, the probability that the load exceeds a certain value 'zl',

$$Pr[Z > zl] = 1.0 - Pr[Z \leq zl] = \int_{zl}^{\infty} f(s)ds \tag{2.16}$$

This is our familiar ELDC. Recall that in the method of cumulants the distribution is given by the Gram-Charlier series expansion,

$$ELDC(z) = \int_{z}^{\infty} f(s)ds = \int_{z}^{\infty} N(s)ds + \sum_{i=3}^{\infty} \frac{\kappa_i}{i!} N^{(i-1)}(z) \tag{2.17}$$

The unit in the loading order 'r' will be committed if the load exceeds the generation of the system up to and including unit (r-1), i.e. if the load exceeds CS_{r-1}

2.2.4 Expected Energy

The expected energy of a unit called into service can be found by either the method of area under the LDC or by the method of energy balance (Stremel et al 1980), i.e.

$$EE_r = UE_{r-1} \text{ - } UE_r$$

The later is illustrated in Fig. 2.9 where 'UEr' is the Unserved energy of the system after unit 'r' with capacity 'Cr' has been loaded. With this unit on-line the system capacity is now 'CSr'. The method of energy balance will be used in this work.

In the case where EE_r represents the hydro block of interest the value of EE_r must be equal to E_{nb}, the available hydro energy of the block. The unserved energy after loading unit 'r' is given by,

$$UE_r = \int_{y}^{\infty} \int_{z}^{\infty} N(s)dsdw + \int_{y}^{\infty} \sum_{i=3}^{\infty} \frac{\kappa_i}{i!} N^{(i-1)}(z) \tag{2.18}$$

where,

$$y = \frac{(CS_r - m_1)}{\sigma} \qquad (Normalized \quad CS_r) \tag{2.19}$$

This can be written as,

$$UE_r = N(y) - y \int_{y}^{\infty} N(s)ds + \sum_{i=3}^{\infty} \frac{\kappa_i}{i!} N^{(i-2)}(y) \tag{2.20}$$

The condition that must be met is that all the hydro energy E_{nb}, must be used, i.e. the hydro is loaded when E_{nb} balances the energy demand. We want the regulating hydro block to be used

as high as possible in the loading order compatible with all the hydro energy being utilized. The higher the position in the loading order of the hydro, the more likely that the largest amount of most expensive thermal plants would be off-loaded. At each successive loading point a test is performed to check the possibility of bringing up the regulating hydro block, by comparing the unserved energy after loading the unit of that particular position in the loading order list, with E_{nb}. Three cases must be distinguished:

$$Case\ 1 : E_{nb} < UE_r \tag{2.21}$$

$$Case\ 2 : E_{nb} = UE_r \tag{2.22}$$

$$Case\ 3 : E_{nb} > UE_r \tag{2.23}$$

Case 1. In this case the hydro block, with energy E_{nb} is not loaded. The following plant in the loading order, 'r+1', is loaded instead. Note that every time the hydro block is not loaded it is pushed up, hence the strategy automaticaly looks for the topmost position for this block. Eventually one or more units or blocks can be withdrawn from the system generation depending on the hydro energy available.
Case 2. In this case the energy demand after loading the '(r-1)-th' unit, EU_{r-1}, is equal to the available hydro energy, hence the hydro plant is loaded. The hydro is placed in position 'r', and the unit formerly in that position is left out of the system. The hydro block in this case is thus used for peak shaving of the load.
Case 3. In this case the available hydro energy shuld be used to de-load the unit in position 'r', to allow optimum discharge of the water. If it is possible to partition the r-th unit into several blocks each with a constant incremental cost, then one should off-load the blocks of the r-th unit starting with the most expensive block until one reaches the point where all the hydro energy is utilized. The remaining blocks are thus loaded after the hydro, if necessary.

When the position of the hydro block in the loading order has to be determined, it is possible that the calculated capacity of the hydro block does not match the rated capacity 'C_R. In this case we define an equivalent capacity C_h, for the hydro block, such that if this capacity was utilized then the amount of hydro energy used in this position matches E_{nb}. Note that $C_h \leq C_R$ Finally, the Loss of Load Probability (LOLP) can be calculated. By definition the ordinate of ELDC at the point CS_T gives the LOLP since loads greater than the total generation of the system cannot be supplied, that is,

$$\text{LOLP} = \text{ELDC(w)}$$

where,

w = standardized CS_T
CS_T = Total system generation capacity

2.2.5 Example

To illustrate the technique discussed above consider the example presente in Stremel et al. (1980) where we had added a hydro block. The power system under consideration has 4 plants. The first one is divided in to two blocks: the lower block is loaded first, i.e. loading order (LO) 1; the peak block is loaded in LO 5, see Fig. 2.10. The other plants are loaded as indicated in Fig. 2.11. The load demand for this problem is an hourly monthly curve (period = T), and is defined by the cumulants presented in Table 15. Furthermore suppose that there is an available hydro energy of 3310.4 T MWh and that a margin of unserved energy of 170 T MWh is acceptable. The problem is to find the best position of the hydro energy block and its equivalent capacity. The rated capacity of the hydro is 5500.0 MW. The units in LO 2, 3 and 4 can be de-rated from 1000.0 MW to a number of levels which include 900 and 857.3 MW. Some results are presented in Table 16. The unserved energy refers to the energy not met after the block/unit of columnn 1 has been loaded.

Table 15. Cumulants of load demand

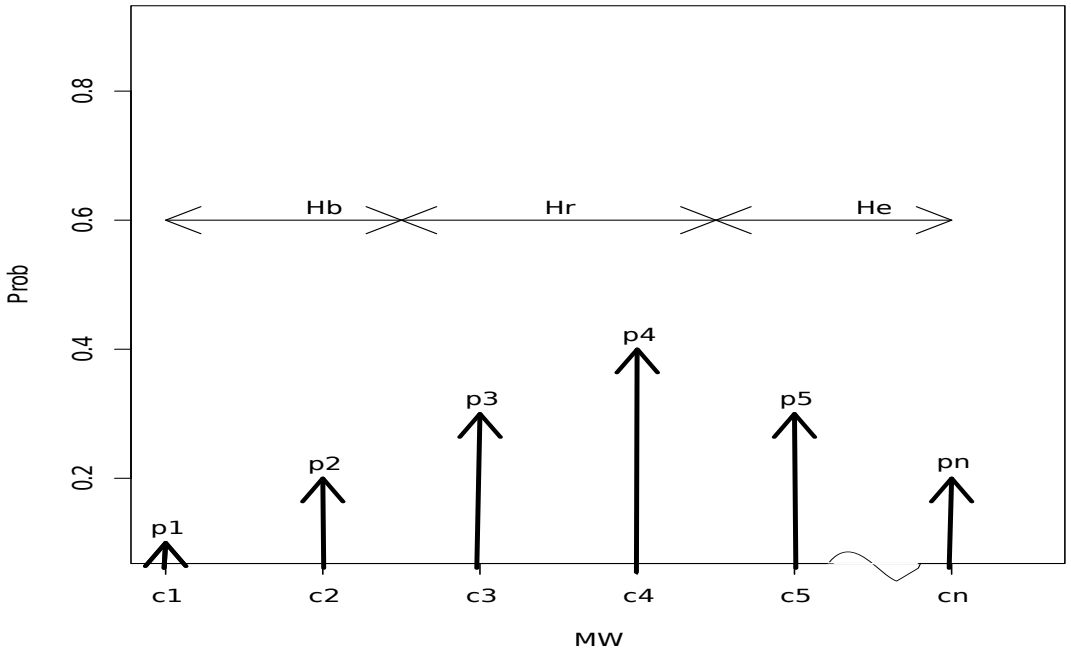

Figure 2.8: Discrete distribution of a hydro plant

Order	Cumulant
1	6000.0
2	40000.0
3	-0.24E+10
4	0.96E+12

Table 16. Expected energy of the first 4 plants

Block/Unit in LO	Expected energy	Cost	Unserved energy
1 (200) MW	179.3 T MWh	A$1434.4 T	5823.3 T MWh
2 (1000) MW	892.94 T MWh	A$8036.5 T	4930.3 T MWh
3 (1000) MW	783.1 T MWh	A$7048.0 T	4147.2 T MWh
4 (1000) MW	666.8 T MWh	A$6668.0 T	3480.4 T MWh

Table 17. Expected energy of the hydro in LO 5

Available hydro	0.331E+04 T MWh
Equivalent capacity	5442.5 MW
Rated capacity of the hydro block	5500.0 MW
System capacity de-loaded	600.0 MW
On-line capacity	8642.5 MW
Total energy generated	0.583E+04 T MWh
Cost of energy	A$0.232E+05
Unserved energy	0.17E+03 T MWh
LOLP	0.171

After unit 4 has been loaded, the unserved energy, accounting for the acceptable margin, has a value of 3310.4 T MW which is exactly the available hydro energy, hence eq. 2.22 holds and we have case 2. The hydro is loaded instead of the next unit (block B2). Then the equivalente capacity is calculated. The result are shown in Table 17.

Case 3: For the more general case suppose that the available hydro energy for the period T is 3400.0 T MWh. The initial results presented in Table 17 do not vary. In this case however, the unserved

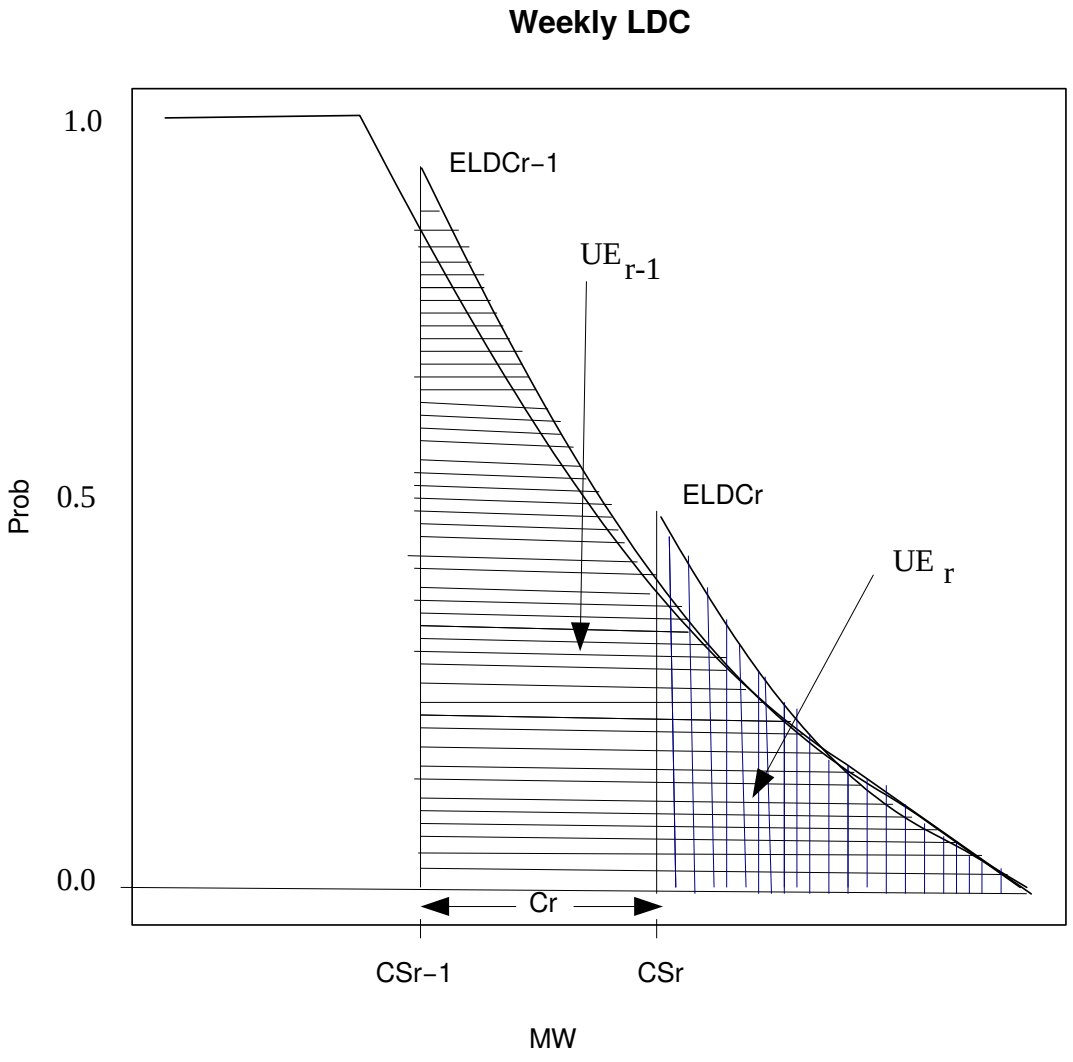

Figure 2.9: Unserved energy

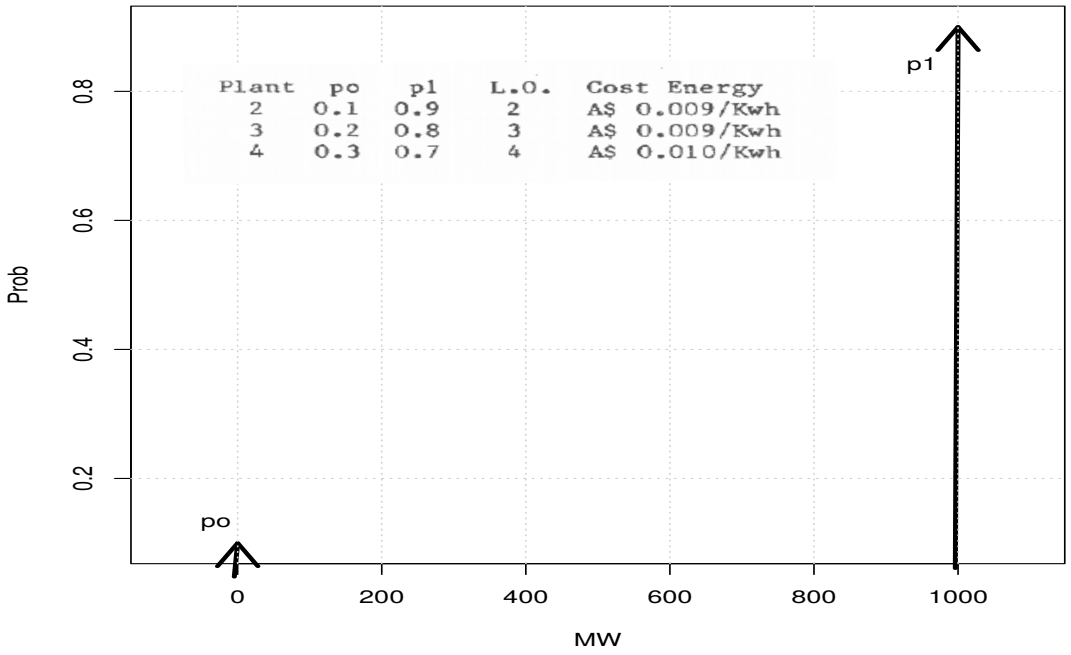

Figure 2.10: Model for the second, third and fourth plants

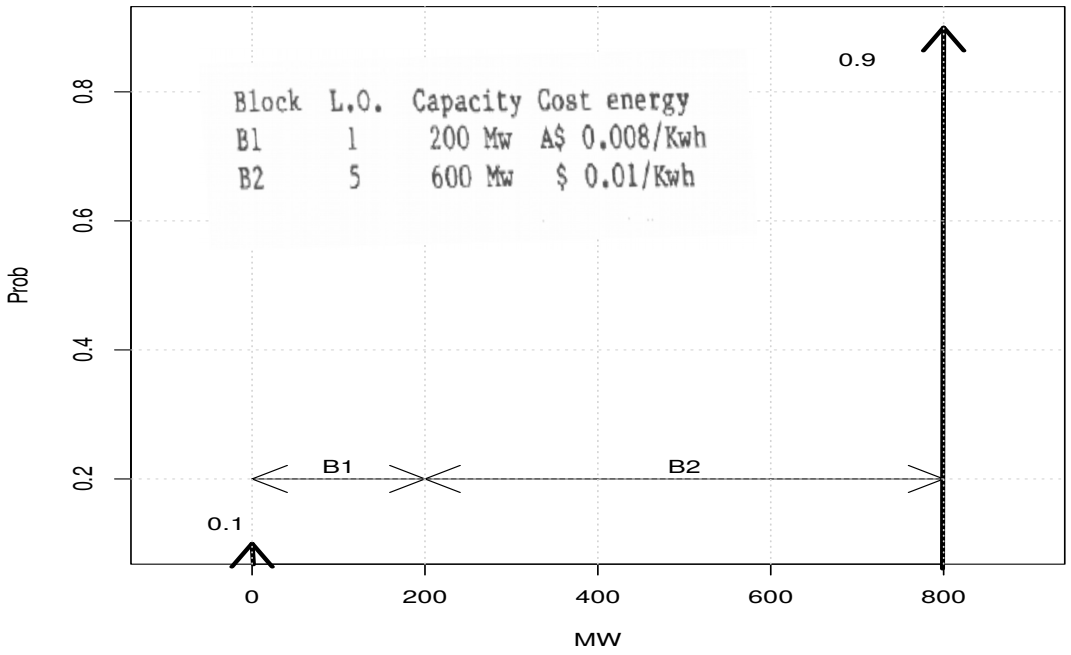

Figure 2.11: Model for the first plant

energy after unit 4 has beel loaded is smaller than the available hydro energy (eq. 2.23). To allow loading of the hydro some part of the unit in LO 4 must be de-loaded. The results presented in Table 18 show that this unit is de-loaded to 857.3 MW, i.e. 14.37% of the unit is de-loaded. To supply the load, a part of the remaining block is loaded after the hydro. The final position of the units is presented in Fig. 2.11.

Table 18. Final position of the units/blocks

Available hydro	0.34E+04 T MWh
Equivalent capacity	5498.2 MW
Rated capacity of the hydro block	5500.0 MW
Unit/block in LO	is de-rated to
4 (1000) MW	857.3 MW
6 (142.7) MW	42.7 MW
System capacity de-loaded	700.0 MW
On-line capacity	8598.2 MW
Total energy generated	0.583E+04 T MWh
Cost of energy	A$0.223E+05
Unserved energy	0.17E+03 T MWh
LOLP	0.171

2.3 References

Sanabria L.A., Dillon T.S. (1986). Stochastic Power Flow using Cumulants and Von Mises functions. Electrical Power Energy Systems. Vol. 8, No. 1, January 1986.

Allan R.N., Al-Shakarchi M.R.G. (1977). Probabilistic Techniques in A.C. Load Flow Analysis. Proc. IEE, Vol. 124, No. 2, Feb. 1977.

Stremel J.P. Jenkins R.T. Babb R.A. Bayless W.D. (1980). Production Costing Using the Cumulant Method of Representing the Equivalent Load Curve. IEEE Trans on PAS. Volume: PAS-99 , Issue: 5 , Sept. 1980.

Finger S. (1979). Electric Power System Production Costing and Reliability Analysis including Hydro-electric, Storage, and Time Dependent Power Plants. MIT Energy Lab. Technical Report MIT-EL-79-006.

Contaxi E. Papachristou D. Contaxis G. (2003). Modeling of Combined Cycle Units in Probabilistic Production Costing Models. Proc. 2003 IEEE Bologna Tech Conference, June 23-26, Bologna, Italy.

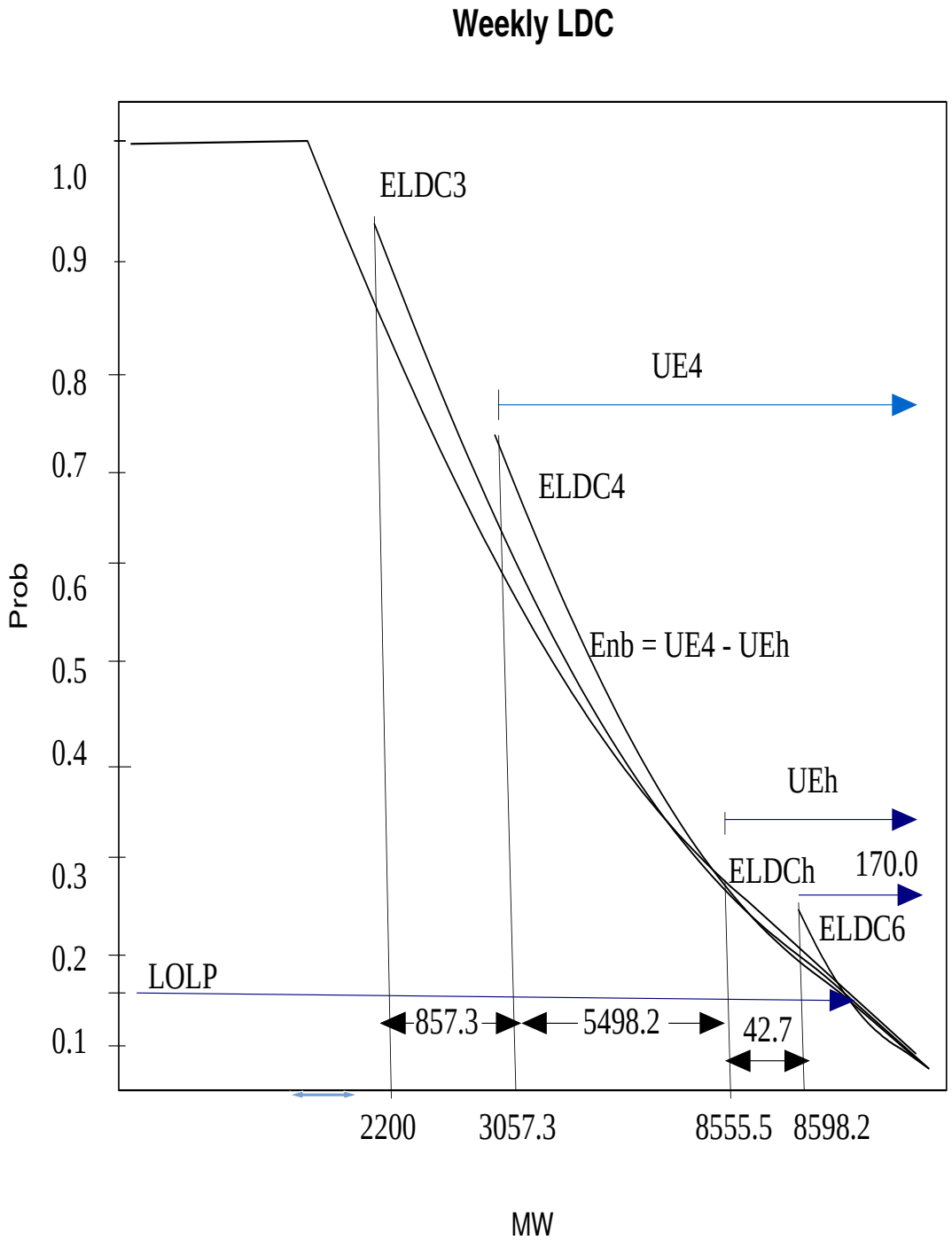

Figure 2.12: Unserved Energy after the units are loaded

Chapter 3

Human behaviour in building fires

This chapter is based on a proyect to simulate fires in buildings developed in the Centre for Environmental Safety and Risk Engineering (CESARE) of Victoria University in Melbourne, Australia. This 5-year project was finalised in 2002.

In this chapter we deal only with the human behaviour part of the project. The Human Behaviour submodel is part of a larger model to quantify the performance of fire-safety systems in Australian buildings. This program, called 'CESARE-Risk', was developed with the support of the Australian Fire Code Reform Centre Ltd. (FCRC). FCRCs aim is to develop a cost-effective, engineering-based approach to fire safety design. CESARE-Risk has been used to reform some aspects of the existing Building Code of Australia.

3.1 Introduction

The CESARE-human behavior submodel (CHBS) aims to simulate people's behaviour when they are involved in a building fire. The model is made up of three components: 1) Response 2) Occupants Evacuation and 3) Fire Brigade and Staff Intervention. Additionally the human behaviour submodel interacts with the CESARE Fire Growth and Smoke Spread program that is used to calculate the fire growth in the apartment of fire origin and smoke spread throughout the building (He 1998).

Due to the uncertainty of human behaviour in fires the CHBS is a probabilistic model. This is because the time-dependent location of occupants in the building during evacuation is treated as a random variable. The probabilistic model has been developed by CESARE researchers based on extensive interviews of people who have experienced actual fire incidents and a review of the existing literature on the subject (Brennan and Horasan 1998).

The methodology used in the CESARE-Risk model consists on setting up a static event tree of fire scenarios describing conditions in a building such as operation of sprinklers, alarms, whether the doors or windows are open, whether is nighttime and so on. These scenarios can run into several thousands each occurring with some probability (Zhao and Beck 1997).

Given the occurrence of a particular scenario the problem is then to calculate the number of fatalities taking into account the time-dependent, non-stationary stochastic nature of fire growth, smoke spread and human behaviour, each of which has infinite number of different realisations. The problem has been solved by calculating the average (expected) outcome over all realisations of a particular scenario. The process is then repeated for each one of the scenarios. The global expected outcome can then be calculated as the average over the various scenarios weighted by the scenario probability. To reduce the problem to manageable proportions, a limited number of representative realisations for each scenario can be selected. Studies carried out at CESARE show that tree realisations can produce approximate results using a fraction of the computer resources as was explained in Section 1.6.1.

In the CESARE-Risk program the Human Behaviour submodel is run 384 times to include different fire conditions, different probabilities that doors and windows in the Apartment of Fire Origin (AFO) are open during the fire, whether the occupants are awake or sleeping, and so on. A description of the CESARE-Risk model is presented in Zhao (1999). The corresponding computer program was developed in C/C++ using a Borland compiler in a Windows environment.

3.2 Response part

In the response part of the CHBS, the model distributes occupants in the apartments according to the percentage of Occupants Groups (OG) in the building nominated by the program user. For residential buildings six basic Occupant Groups (OG) have been identified, namely,

- OG 1 and 2: lone person less and older than 70 years old respectively.
- OG 3: couple and child.
- OG 4: two related or unrelated occupants.
- OG 5: lone person, drugged or intoxicated.
- OG 6: two persons older than 70, one of them with a handicap.

A discussion of the methodology used to select these basic Occupant Groups is given in Brennan (1999). The number of occupants in each OG can be varied by the user to match a desired number of occupants in the building. The program recognises the fact that a lot of people go to work during the daytime so if the fire happens during the day the number of occupants in the building is reduced by a Daytime Occupancy Factor.

In a case of a fire occurrence it is assumed that the occupants start acting after they have received and recognised cues indicating that something unusual is happening. The program can generate four kinds of cues: 1) alarm 2) smoke 3) different type of warnings from other occupants evacuating and 4) sound of window breaking glass. A brief description of each one of these cues follows:

- Alarms: currently the program can deal with nine types of alarms, including different apartment or smoke alarms, different corridor or building alarms, Early Warning Integrated Systems (EWIS) and occupant-activated or break-glass alarm.
- Smoke: smoke in the building apartments is classified based on peoples visibility as No Smoke (visibility greater than 15 m), Light Smoke (visibility between 8 and 15 m), Medium Smoke (visibility between 1 and 8 m) and Heavy Smoke (visibility less than 1 m); see He and Brennan (1998). A smoke cue in an apartment is generated by the program when smoke reaches Light Smoke conditions.
- Warnings from other occupants: the program can generate 3 kinds of warning cues: Direct or indirect warning to occupants of apartments of non-fire origin (ANFO) from occupants of the apartment of fire origin (AFO). Direct or indirect warning to ANFO occupants from other ANFO occupants in the same level and Indirect warning from occupants located in the stairways.
- Sound of window breaking glass: this cue is produced in the program when the average temperature in the room of fire origin (RFO) reaches 300°C.

The number of occupants who recognise a cue is given by the probability of recognition of that particular cue P_{rec}. Those occupants who recognise a cue have one of three options: they can directly evacuate from the apartment, they can investigate conditions in the corridor before they take a decision or they can decide to remain in the apartment and wait for more cues.

The number of occupants in the first two possible options is proportional to a Probability of Direct Evacuation P_{dir_evac} and a Probability of Investigation P_{inv} respectively. These probabilities are given by,

$$\begin{aligned} P_{dir\_evac} &= P_{init\_act,evc} * p \\ P_{inv} &= P_{init\_act,inv} * P_{inv\_act} * p \end{aligned} \tag{3.1}$$

where, the subindex 'evc' refers to those who evacuate; 'inv' refers to those who investigate.
P_{init_act} is the probability that the occupants start moving given that the cue has been recognised.

P_{inv_act} is a probability applied only to those who investigate. It splits these occupants between those who subsequently evacuate and those who subsequently decide to remain in the apartment after the investigation has been completed. The variable p is given in Fig. 3.2 and will be explained below.

Occupants who choose to remain in their apartments may wait for more cues before deciding to evacuate. Fig. 3.1 shows a diagram of the Response Submodel activities. No/light in the figure

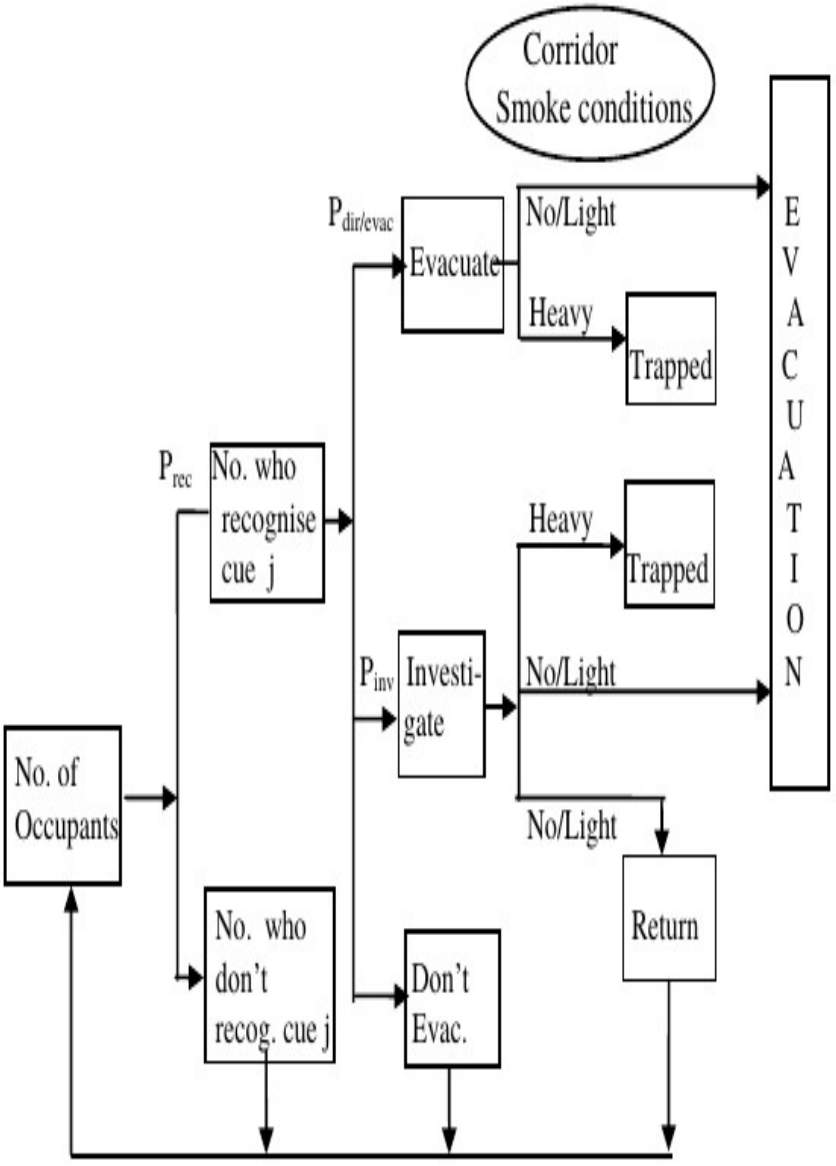

Figure 3.1: Flow of activities in the Response submodel

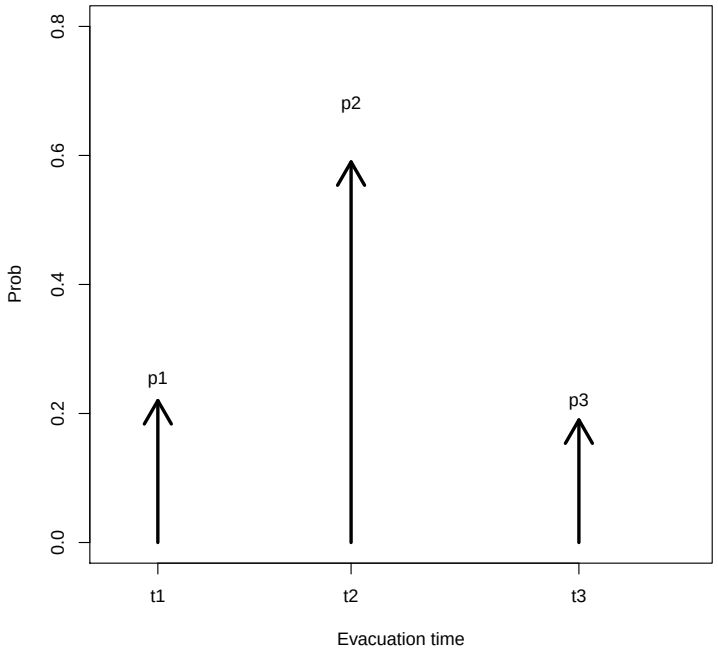

Figure 3.2: Three-point discrete distribution for time to start evacuation

refers to the smoke conditions in the corridor, that is no smoke or only light smoke. Heavy means heavy smoke in the corridor.

Program users can also consider balconies in the apartments. Occupants who are awake and have not evacuated after all cues have occurred, can move to the balconies and wait there safely to be rescued by the fire brigade intervention. Program users can also allow occupants of the first and second floors to evacuate through the windows.

As mentioned before the CHBS is based on a probabilistic model. CESARE researchers have found that the main uncertainty in evacuation is the time at which occupants leave their flats and start moving into the corridor. So this input variable is represented in the program by a random variable. As mentioned in Section 1.6.1 a practical way to deal with this rv is to replace its distribution function by a three-point equivalent discrete distribution as shown in Fig. 3.2.

Based on the three-point distribution, three realisations for evacuation time are considered in the solution. The result of the simulation in terms of expected values is a combination of the results of the three realisations. In practice the program splits the number of occupants evacuating in to 3 subgroups, these subgroups evacuate at times t1, t2 and t3. The number of occupants in each subgroup is proportional to p1, p2 and p3 of Fig. 3.2. The times and numbers of occupants evacuating their apartments are passed to the next submodel.

3.3 Evacuation part

This submodel is used to calculate the evacuation of occupants, once they have decided to leave their apartments; this is represented by the last block in Fig. 3.1. It is assumed that the occupants initially move into the corridor and later start walking downstairs to the building exits at the ground level. The program calculates the time taken for each OG to move to each enclosure. Congestion is an important part of this submodel. The program recognises two kinds of congestion: door congestion and enclosure congestion. The first one is due to the door width, the second one is due to the enclosure capacity. If there is congestion, the program increases the evacuation time accordingly. The program also recognises the fact that not all occupants can evacuate at the same speed, so the evacuation speed of some OGs is affected by a Speed Reduction Factor.

The evacuation submodel calls the Smoke Spread function to calculate the amount of heat, smoke and gases the occupants are exposed to, at every time step. Based on this amount the program

classifies the occupants as mobile or non-mobile. Non-mobile occupants are made up of fatalities, incapacitated and trapped occupants. Trapped occupants are occupants who are all right but cannot evacuate due to heavy smoke in the corridor. Disabled occupants are also classified as non-mobile occupants regardless of the smoke conditions. It is assumed that the non-mobile occupants cannot move by themselves; they have to be helped by the Fire Brigade.

To summarise, the evacuation submodel calculates the number, location and status of all occupants evacuating the building at each time step.

3.4 Fire Brigade Intervention part

This submodel deals with the Fire Brigade (FB) intervention. In some buildings like nursing homes, hotels and so on, there can also be staff intervention. The FB (or staff) intervention submodel is used to modify the number of occupants who are deemed to be incapacitated, fatalities and evacuees as a result of the fire. The FB is split into 3 subgroups: search rescue, fire fighting and officers in charge. The staff intervention has only search rescue subgroup. The main activities of the FB are: arrival to the site, search, rescue of non-mobile occupants, the provision of evacuation instructions to mobile occupants, fire fighting and the call for more resources if necessary (this last activity is carried out by the Officer in Charge).

The FB intervention submodel is also based on a probabilistic model developed by CESARE and FB researchers Zhao, Beck and Kurban (1998). The random variable in this case is the FB arrival time, because the FB truck speed is a random variable. The program handles this random variable by calculating a three-point discrete distribution of truck speed from the distribution function of the original random variable. From this discrete distribution a three-point distribution of arrival time is calculated, as explained before. Based on the three-point distribution, three independent problems or realisations are solved with three FB arrival times. The final result is a combination of the individual results affected by the realisation probability.

The program recognises only two outcomes for the fire fighting activity: fire under control if the water resources of the Fire Brigade are greater than the heat release rate, or fire not under control otherwise. The heat release rate is calculated by the Fire Growth submodel. If the fire is under control the program calculates a ramp function to slowly reduce the impact of the fire, smoke and gases on the occupants. A typical ramp function is shown in Fig.3.3. In this figure 'tfuc' is the time at which the FB has the fire under control, 'text' is the time at which the fire has been extinguished and 'tsp' is the time at which the smoke has been sparsed. The default value for 'text' is 7.5 min. The default value for 'tsp' depends on the fire type: for flashover fires 'tsp' is 60 min. For flaming fires 'tsp' is 30 min. For smouldering fires 'tsp' is 5 minutes.

3.5 Example

To illustrate the main characteristics of the model discussed above, a small example case will be solved. Consider a fire in a 3-storey residential building with 6 apartments per floor, one corridor, two stairways and with balconies in the apartments. The fire originates in the first floor, during the night. A single person less than 71 years old (OG 1) is in the Apartment of Fire Origin (AFO) at the time of the fire. This occupant is located in a room other than the Room of Fire Origin, called the Room of Non-Fire Origin (RNFO). The building is fitted with building alarms with local detectors inside the apartments. The dimensions of the building floor are presented in Table 3.1.

Table 3.1. Dimensions of the building floors (m)

Element	Length	Width	Door width
Apartments	25.5	6	0.9
Corridors	30	3	1.1
Stairways 1 and 2	9	1.65	1.1

Note that Door width in the case of corridors and stairways is actually the door between corridor and stairways. As explained in Section 3.1 the program distributes Occupant Groups (OG) in the building, based on user-requested percentages. Table 3.2 shows the distribution of OGs for the example problem.

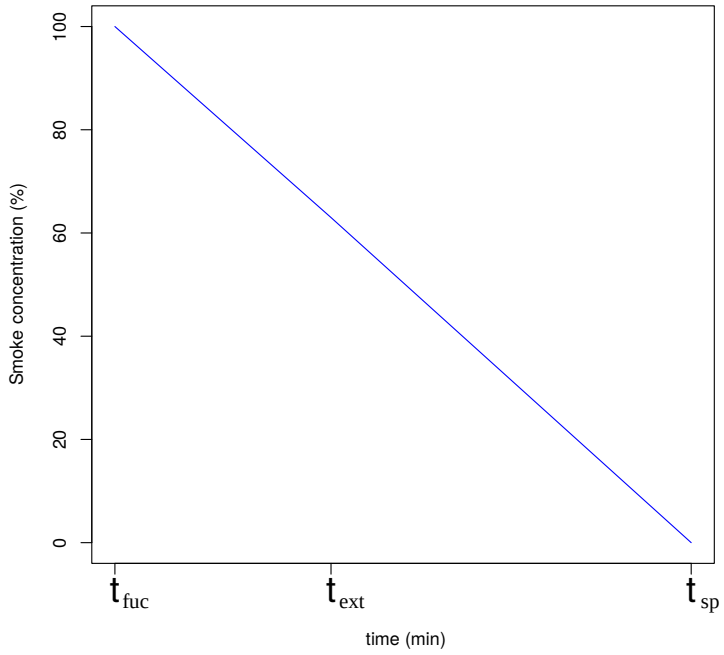

Figure 3.3: Ramp function to reduce impact of smoke on occupants

Table 3.2. Occupant distribution in apartments

Floor	Apartment	OG type	No. of persons	Total persons per floor
1	AFO	1	1	-
1	2	1	1	-
1	3	2	1	-
1	4	3	3	-
1	5	4	2	-
1	6	5	1	9
2	1	1	1	-
2	2	3	3	-
2	3	4	2	-
2	4	1	1	-
2	5	4	2	-
2	6	4	2	11
3	1	1	1	-
3	2	2	1	-
3	3	3	3	-
3	4	4	2	-
3	5	1	1	-
3	6	4	2	10
Total	-	-	-	30

A discussion of the probabilities used in Equations 3.1 and their collection methodology is provided in Brennan and Horasan (1998). The values for the three-point discrete distributions used in the example are presented in Table 3.3, see also Fig. 3.2.

Table 3.3. Values for three-point discrete distributions.

Direct/evacuation			
Probabilities p1, p2 and p3	0.22	0.59	0.19
Times to start evac t1, t2 and t3 (sec)	11.7	62	121.3
Investigate			
Probabilities p1, p2 and p3	0.47	0.19	0.33
Times to start evac t1, t2 and t3 (sec)	100.4	368.9	1266.5

As explained before the Human Behaviour subprogram is called 384 times by the CESARE-Risk program, the example presented here corresponds to scenario 28. The conditions for this scenario are presented in Table 3.4. The program solves the problem in time steps, for this problem 1 time step = 5 seconds.

Table 3.4. Conditions for scenario 28.

Fire type	Flashover
AFO door	Open
Stair door	Open
Occupants status	Sleeping
RFO door	Open
RFO window	Closed
Fire spread to RNFO	Yes

Once the occupants have been distributed in to the building the program starts the fire in the RFO and passes the information to the smoke spread model. The CHBS interacts with these submodules and starts producing the cues one at the time. All the cues produced by the program in this particular scenario are presented in Table 3.5. Notice that the building is fitted with building alarms with smoke detectors inside the apartments.

Table 3.5. Cues produced in this example case.

Cue	Location	At time (min)
Smoke	AFO	1.8
Alarm	AFO	2.1
Alarm	1st floor	3.1
Alarm	2nd floor	3.1
Alarm	3rd floor	3.1
Sound of breaking glass	AFO	3.5
Warning	1st floor	4.3
Warning	1st floor	4.7
Warning	1st floor	6.0
Warning	2nd floor	6.3
Warning	2nd floor	8.1
Warning	2nd floor	8.1
Smoke	3rd floor	10.8
Smoke	2nd floor	11.8
Smoke	1st floor	18.5
Untenable conditions in AFO		5.75

The first calculation carried out by the program is the number of occupants who recognise the first cue. Based on this calculation, the expected number of occupants who directly evacuate and the expected number of occupants who investigate before taking a decision is calculated using Equation 3.1.

The next cue produced by the program is the alarm cue. In the program the AFO detector detects the smoke and operates inside the apartment at time step 25 or 125 seconds. The building alarm sounds at 37 time steps or 3.1 minutes. Again the program calculates the number of occupants who recognise this cue, this number is then split into occupants who directly evacuate and occupants who investigate before taking a decision, as explained before.

3.6 Results

The results of the evacuation process for this particular scenario for a simulation of 50 minutes, are presented in Table 6. These results do not include the effect of Fire Brigade intervention. Notice that all times are from the start of the fire.

Table 3.6. Results without FB Intervention.

Total No. of occupants	30
No. of evacuees	1.78
Time to complete evacuation	5.83 min
No. of fatalities in AFO	0.017
No. of fatalities in Apartments of LFO	0.0
No. of fatalities in Corridors of LFO	1.78
No. of occupants in Apartments of LFO	2.5
No. of occupants in balconies (LFO)	4.71
No. of fatalities in Apart. of 2nd floor	0.397
No. of fatalities in stairs of 2nd floor	1.19
No. of occupants in balconies 2nd floor	9.42
No. of fatalities in Apart. of 3rd floor	0.61
No. of fatalities in stairs of 3rd floor	1.05
No. of occupants in balconies 3rd floor	8.35
Total fatalities	5.05
Place where 1st fatality occurred	AFO
Time at which 1st fatality occurred	5.75

For the FB intervention let us assume that the initial resources of the FB, for each one of the 3 realisations, are as presented in Table 3.7, taken from Zhao et al (1998).

Table 3.7. Fire Brigade initial resources for all realisations.

Trucks	Total fire fighters	Search and Rescue group	Fire fight group	Officers in Charge
4	12	4	4	4

The first calculation carried out by the FB submodel is the three-point distribution of time for the FB to start the fire fighting and search & rescue activities, the results are presented in Table 3.8. Time to start is from the start of the fire.

Table 3.8. Distribution of Time to Start.

Realisation	Time to start	Probability
1	7 min	0.19
2	8.58 min	0.61
3	13.3 min	0.2

Based on Tables 3.6 and 3.7 the FB Submodel calculates the expected number of occupants rescued by the FB and their status, as presented in Table 3.9. All times are from the start of the fire. Note that the firefighters faced untenable conditions whilst searching the AFO and had to abort the search and rescue operation in this room in all realisations (Real.). The same happens during search in LFO in Real. 2, this is marked in the table with 'Ab.' for 'Operation aborted'.

Table 3.9. Results of FB intervention.

Activity	Real. 1	Real. 2	Real. 3	Total
Probability	0.19	0.61	0.2	1.0
Start activities (min)	7.75	9.33	14.1	
Search in AFO	Ab.	Ab.	Ab.	
Time to complete search in LFO Apt. (min)	14	Ab.	20.3	
Fatalities recovered from LFO corridor	0.17	-	1.78	0.39
Disabled occupants rescued from LFO	1	-	1	1
Time to complete activity in LFO (min)	17.8	-	29.3	
Time to complete search in 2nd floor (min)	24.5	22.2	36.1	
Fatalities recovered from 2nd floor apart.	0.0	0.0	0.4	0.08
Fatalities recovered from 2nd floor stairs	1.19	1.19	1.19	1.19
Time to complete rescue in 2nd floor (min)	28.6	26.3	41.6	
Time to complete search in 3rd floor (min)	35.5	33.2	48.5	
Fatalities recovered from 3rd floor stairs	1.04	1.04	1.04	1.04
Fatalities recovered from 3rd floor apart.	0.36	0.61	0.61	0.56
Time to complete activity in 3rd floor (min)	40.33	38.8	54.2	
Fire under control at (min)	13.5	40.1	36.5	

The results of the simulation for scenario 28 with FB intervention are presented in Table 3.10. This table can be compared with Table 3.6 to see that the FB intervention has reduced the number of

fatalities by more than 50%. Note that the FB has instructed the occupants in the apartments to evacuate including occupants in the balconies, the number of occupants who have evacuated is now 7.98. As expected these occupants need more time to complete evacuation.

Table 3.10. Results with FB Intervention.

Total No. of occupants	30
No. of evacuees	7.98
Time to complete evacuation	21.4 min
No. of fatalities in AFO	0.017
No. of fatalities in Apartments of LFO	0.0
No. of fatalities in Corridors of LFO	0.0
No. of occupants in Apartments of LFO	0.0
No. of occupants in balconies (LFO)	0.0
No. of fatalities in Apart. of 2nd floor	0.0
No. of fatalities in stairs of 2nd floor	1.19
No. of occupants in balconies 2nd floor	0.0
No. of fatalities in Apart. of 3rd floor	0.364
No. of fatalities in stairs of 3rd floor	1.04
No. of occupants in balconies 3rd floor	0.0
Total fatalities	2.25
Place where 1st fatality occurred	AFO
Time at which 1st fatality occurred	5.75

The results of all the 384 cases considered in CESARE-Risk after all runs have been completed are presented in Table 3.11. Case I in the table refers to apartments with alarms while case II refers to building alarms. The real number of fatalities in these tables seem extrange at first, they were the cause of much hilarity within the CESARE-Risk team. It must be remember that these results are representative of 1000 fires in residential buildings considering a large number of different characteristics of building and occupants. Hence occupants and fatalities in the tables above should be multipplied by 1000, i.e. the number of expected fatalities in the apartments of the 3rd floor in Table 3.10 is 364 occupants per 1000 fires.

Table 3.11. Results of CESARE-Risk full run.

Location of fire	Percentage	Predicted fatalities Case I	Predicted fatalities Case II
Kitchen	32%	8.6	5.1
Bedroom	42%	45.8	23.1
Lounge	26%	39.4	20.3
Weighted average		21.1	11.2

Chapter 4

Constrained Stochastic Optimisation

In this chapter we discuss a technique for solution of general constrained stochastic optimisation problems. The central characteristic of these problems is that some or all system parameters are not known with certainty. To capture the uncertainty, random variables have to be used to model these parameters. In some design problems the solution has to fit within certain limits giving rise to the classical constrained stochastic optimisation problem. These problems are difficult to solve using conventional techniques like linear or dynamic programming. In this chapter we use a relatively new technique termed Genetic Algorithms (GA). The limitation of the current solution of stochasting problems using GA algorithms is that they require a very large number of sample points taken over the random distributions to capture the random nature of some of the parameters and hence produce meaningful results. Even with todays computers, realistic-size engineering problems can be difficult to tackle due to limitations in computer resources. In this chapter we present a computer-efficient method to solve these types of problems based on the three-point distribution technique introduced in Section 1.6.1. To illustrate the technique an Intranet server design will be optimised for reliability keeping the server cost under a given value.

4.1 Introduction

In many engineering systems it is important to capture the uncertainties inherent in the system. To do this it is necessary to model the problem using random variables however modelling the problem using these types of variables increases solution complexity. This has given rise to a series of techniques known collectively as stochastic modelling. The aim of these techniques is to find ways to solve the problem. In some cases it is possible to simplify the problem by reducing it to a system of linear equations with random variables and try to find an analytical solution, usually limiting the kind of random variables which can be used. In other cases this is not possible and engineers have developed computer simulations. The best known of them is the Monte Carlo simulation. Stochastic optimisation problems are particularly difficult problems to solve because of the competing needs of modelling the variables with probabilistic functions and at the same time searching large spaces looking for the global optima.

4.2 Constrained Stochastic Optimisation (CSO) problems

The general formulation of the Constrained Stochastic Optimisation problem is:

Maximise f(x)
Subjected to

$$
\begin{aligned}
& C_F \leq C_{MAX} \\
g_i(x) = 0 \quad & for \quad i = 1, 2, 3 \cdots n_1 \\
h_j(x) \leq 0 \quad & for \quad j = 1, 2, 3 \cdots n_2
\end{aligned}
\tag{4.1}
$$

Where some or all x are random variables. C_F is the cost associated with a particular point of the f(x) space and C_{MAX} is the maximum permissible cost.

In the conventional solution of the problem the random variables are replaced by their expected values given rise to a multi-variate stochastic integration problem. These kinds of problems can be solved using Monte Carlo simulation. The problem with Monte Carlo simulation is that large amounts of sample points have to be used for each of thousands of runs imposing huge demands in computer resources.

In this sectioni we present a more efficient technique to solve constrained stochastic optimisation problems based on the concept of the three-point equivalent distribution. The three-point equivalent distribution coupled with optimisation via genetic algorithms provides a computer-efficient solution to the problem.

The use of the three-point distribution greatly simplifies the stochastic optimisation problem because we need to consider only three points from a probabilistic distribution. Furthermore the three-point distribution is an equivalent distribution, no inaccuracies are introduced in the solution. The algorithm renders the probability distribution of solution points. In some applications such as fire protection engineering, the parameter of importance is the expected value of the process. It indicates the statistical tendency of a large number of observed events. In other problems the value of interest is the maximum possible value of all events as in the reliability optimisation problems.

4.3 Genetic Algorithms in Optimisation

Optimisation via Genetic Algorithms was a topic introduced by Holland in 1975 and popularised by Goldberg in the 80s (Golberg 1989). A large amount of engineering problems have been solved using Genetic Algorithms, see for instance Gen and Runwei (1997), Pham and Karaboga (2000). GA applications in reliability optimisation were pioneered by Coit and Smith (1996), Painton and Campbell (1995).

The technique is a systematic random search of solution points over the space of feasible and unfeasible solutions. The strength of the method relies on the fact that the actual parameters of the function to be optimised are used in the search so there is no need to calculate auxiliary information like gradients or hill climbing directions. GA evaluates a large amount of surface points simultaneously thus the probability of being confused by a local or false solution is reduced. The method produces acceptable results even for problems with difficult shapes and local minima. The systematic search is guided by a process similar to the way living organisms survive tough environments. It uses reproduction, crossover and mutation.

The first step in the optimisation of engineering systems via Genetic Algorithms is to encode the system variables in to a string of values called a chromosome. Most GA techniques use binary values for encoding the variables. The chromosomes fitness is the function value for a given set of parameters. The first chromosome can be generated at random, its variables decoded to decimal values and the function to be optimised (fitness) calculated: this is the first point in the optimisation process. The total number of chromosomes used in the process is chosen arbitrarily and is called a population. The chromosomes fitness can be improved by crossover with a group of carefully selected chromosomes.

In the selection process all chromosomes' fitness is ranked in proportion to its value. The fittest chromosomes appear more times in the selection process, this is the so-called weighted roulette wheel. Chromosomes in the roulette wheel are crossed by coping the head of one with the tail of another one starting at a random point of the chromosome length. Two children are born of such crossing, normally the children replace the parents in the next generation of chromosome population. Chromosomes are also subjected to mutation. In mutation, a bit of the chromosome length selected at random is flipped to the binary complementary. Mutation is there to add diversity to the population and protect against premature convergence.

The number of crossovers and mutations are determined using a probability of crossover (pc) and a probability of mutation (pm). If the probability of mutation is, say 0.01, we expect that on average 1% of the total number of bits in the population will suffer mutation in each generation.

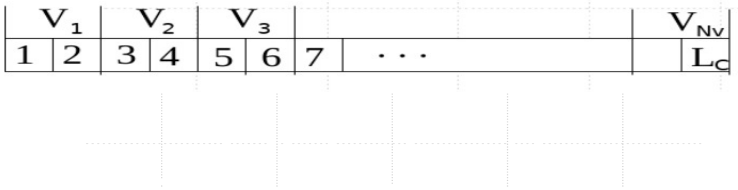

Figure 4.1: A chromosome of a population of n chromosomes.

4.4 Solution of the Constrained Stochastic Optimisation Problem

The first step in the solution of the (CSOP) using the equivalent three-point distribution and a GA algorithm, is to calculate the equivalent distribution for each one of the random variables. If the distribution is known, the statistical moments can easily be found as discusse in Section 1.4. If the distribution is not known but there are historical records, the variables moments can be calculated from the records. Once the moments are known, the three-point equivalent distribution can be calculated as explained in Section 1.6.1.

The second step is to encode the variables into a chromosome. Assuming that all variables are represented by a three-point distribution, they can take only three values: 1, 2 and 3. So only two bits are needed to encode a variable because in binary numbers 3 is represented by the string 1-1. To illustrate the process suppose that the encoded variable 'vi' takes the value 1-0 (which is 2 in decimal numbers) the program selects the second point of the three-point distribution as the value for variable vi.

The total chromosome length is given by,

$$L_C = 2 * N_Y \tag{4.2}$$

Where N_Y is the number of variables to be encoded

Fig. 4.1 shows a set of variables encoded into a chromosome of length L_C Once all variables have been decoded to their corresponding decimal values it is possible to calculate the chromosome fitness. This fitness is the function evaluated using the decimal values,

$$fitness = f(x_1, x_2 \cdots x_v,) \tag{4.3}$$

In some applications, as in reliability studies, f(x) is system failure rate, so for optimisation problems the complementary is used, ie,

$$fitness = 1.0 - f(x_i) \tag{4.4}$$

Each chromosomes fitness has a cost associated with it and a probability of that particular solution happening.

The fitness probability is calculated using eq. (4.5),

$$Prob_i = \prod Pr(\lambda_i) \qquad for \qquad i = 1, 2, 3 \cdots Nv \tag{4.5}$$

Where
λ is a point of the three-point equivalent distribution
$Pr(\lambda)$ is the corresponding probability of λ happening.

And the fitness cost C_F is calculated using eq. (4.6),

$$C_F = \sum Cost_i \qquad for \qquad i = 1, 2, 3 \cdots Nv \tag{4.6}$$

A chromosome is feasible if,

$$C_F \leq C_M \tag{4.7}$$

Where C_M is the maximum permissible cost.

Using the three-point distribution and GA algorithms it is possible to reduce the CSOP to a more manageable problem. The technique will be used to find the optimum design for an Intranet server. A brief description of the server is presented next.

4.5 The e-Hub

The Computer Sciences and Computer Engineering Department, La Trobe University, is working on the development of a highly efficient Web system for a virtual logistics provider as explained in Soh and Sattar (2002).

A virtual logistics provider congregates a number of companies or partners into strategic alliances. Each partner is itself a logistics provider, but they recognise that there are advantages in sharing warehousing and transportation facilities to achieve easier supply over a widely geographically distributed area of operation. This class of consortia has special needs for inter-organisational information exchange and communication hence they can have one or more Intranet servers or e-Hubs that provide points of entry into the full set of facilities. The critical feature of the e-Hub is its reliability because the performance of the e-Hub determines the overall performance of the collaborative system.

4.5.1 Basic e-Hub architecture

For the optimisation problem we start with a basic architecture and work out methods to improve the design. Fig. 4.2 shows a schematic representation of the basic e-Hub system. The system is made up of five processing subgroups: Load Balancer, Web Interface, Security Manager, Global Database and Dispatcher. Each one of these subgroups performs a specific function for the e-Hub and has processing redundancies to increase overall reliability. The system performs its function if a user can access it, browse through the different products the warehouse offers and make a request. The request is, then processed by the Dispatcher and an order to carry out some action follows. This order is considered the system output. It is assumed that at each step the system has enough data checking procedures to guarantee data integrity (Soh and Sattar 2002). In this section we are concerned only with system reliability, that is, our interest is to guarantee a flow of data from input to output.

4.5.2 Reliability Analysis

For reliability analysis we model the e-Hub as a fault-tolerant system. Fig. 4.3 shows a fault-tolerant model of the system. The reliability analysis of the fault-tolerant model was presented in Sanabria et al (2002). As explained in that paper a number of techniques were considered for reliability analysis and the Cut-Set technique was selected for its effectiveness in the solution of fault-tolerant systems made up of series/parallel elements.

In the Cut-Set (CS) technique it is necessary to list all minimal Cut-Sets between input and output. A Cut-Set is a set of components whose failure results in system failure. A minimal Cut-Set is a cut where the set remaining after the removal of any of its elements is no longer a cut (Jasmon and Kai 1985).

A CS made up of one element is termed a first order CS. Cut Sets made up of two elements as are called a second order CS and so on. By inspection it is possible to enumerate all CS of Fig. 4.3 up to third order (higher order CS contribute very little to the final result):

1st order CS: 9
2nd order CS: 1, 2; 10, 11
3rd order CS: 3, 4, 5; 6, 7, 8.

For reliability analysis Fig. 4.3 can be reduced to a system of elements in series. To do this reduction, all second and higher order CS are first reduced to an equivalent first order CS. Reliability calculation of a series system is simply the addition of component reliability (Sharma 1976).

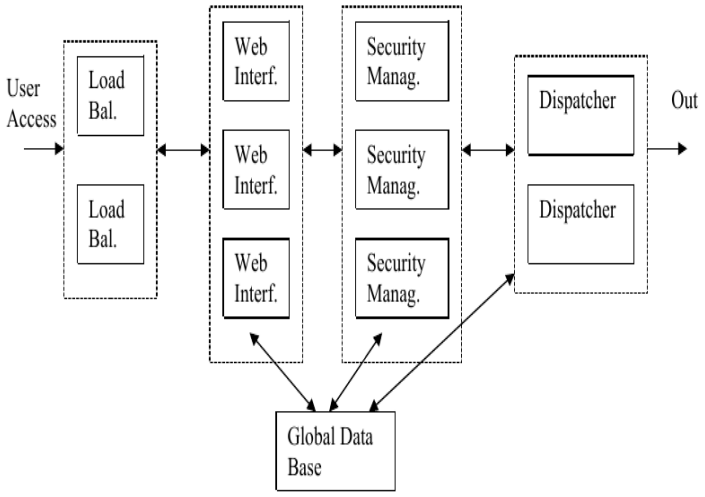

Figure 4.2: Schematic representation of the e-Hub system.

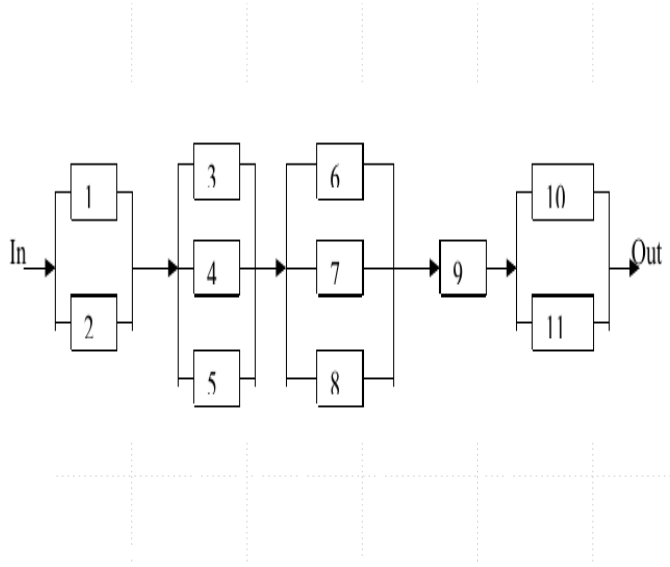

Figure 4.3: Fault-tolerant model of the e-Hub system.

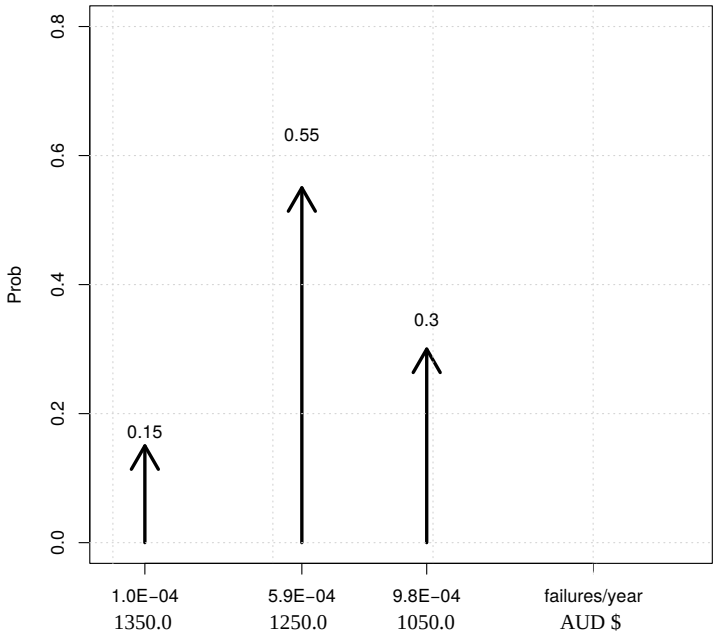

Figure 4.4: Three-point distribution of failure rate and corresponding cost.

4.5.3 Probabilistic Modelling of Failure Rate

Researchers have recognised the need to model the component failure rate of electronic equipment with a random variable, see for instance Gen and Runwei (1997), Painton and Campbell (1995). The rationale for this is that each processor has its own pattern of failure which varies over a wide range. The pattern of failure can be modelled by a random variable using historical data of processor failure. In some cases data can be generated by a computer simulation of processor behaviour. In other cases the trend of failure rate is known and a standard deviation around this trend can be calculated based on previous experience with similar units. In all cases modelling failure rate using random variables produces more realistic results allowing the design engineer a better judgement of system behaviour.

The second reason for using random variables in the reliability analysis is that there is a cost associated with different levels of component reliability. Design engineers should aim to spend their money in the areas where it is more effective. A random variable allows design engineers to select within a wide range of component reliability costs and look for a combination of very reliable and no so reliable components to maximise system reliability at minimum cost.

To sumarise we want to find the most reliable configuration for the e-Hub. The problems random variables are component reliability (it is assumed that identical components are used in a subgroup). The deterministic variables are the number of redundancies in the subgroup. The overall system configuration must not exceed a budget of C_M dollars.

4.5.4 Results

To Illustrate the method lets suppose that historical records for the failure rate of some type of units to be used as the Load Balancer shown in Fig. 4.2, are available. From the records the central moments for the distribution of failure rate can be evaluated and an equivalent three-point distribution can be calculated using eq. (1.33). Further lets assume that the cost of those units varies between A$1050 and A$1350 with mean at A$1250, the more reliable units being more expensive. The equivalent distribution including the cost variation is presented in Fig. 4.4.

The maximum number of redundancies which can be used in any subgroup is 7 and the maximum cost of the e-Hub server must not exceed A$8000.0.

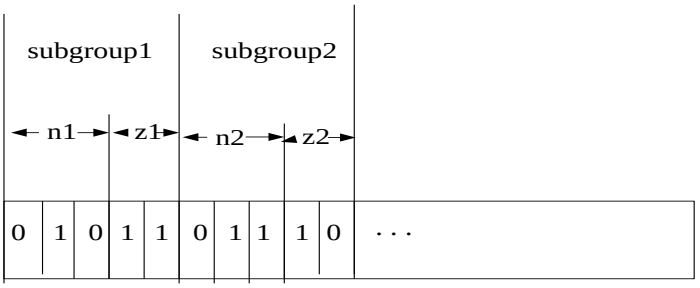

Figure 4.5: Pair of values (n,z) for subgroups 1 and 2 encoded in chromosome 1.

The CS technique requires also repair time, maintenance rate and maintenance time. Some or all of this data can also be given in the form of random variables, so more variables would need to be encoded. Just for simplicity let us assume that these variables are deterministic and that their values are as presented in Table 4.1.

Table 4.1. Repair and maintenance for e-Hub system

Subgroup	Repair time	Maint. rate Outages/year	Maint. time Hours
1	0.2	0.001	0.01
2	0.5	0.0001	0.001
3	0.5	0.00001	0.0001
4	0.5	0.00001	0.0001
5	0.5	0.00001	0.0001

Fig. 4.5 presents the strings to encode the pair of variables n (number of redundancies required) and z (a point of the distribution of Fig. 3.4). In the figure the GA algorithm has selected n1 = 2 (two redundancies) and z1 = 3 (failure rate = 9.8x10-4) for subgroup 1 (Load Balancer) and n2 = 3 and z2 = 2 (failure rate = 5.9x10-4) for subgroup 2 (Web Interface). Table 4.2 presents the values of the whole chromosome. Note: as explained in Painton and Campbell (1995) the CS technique produces overall outage rates per 100 years. Chromosome fitness is calculated as 1.0 - overall outage rate.

Table 4.2. Values of first chromosome

Subgroup	n	z	λ	Fitness	Prob	Cost
1	2	3	9.8E-04	-	0.30	2*1050.0
2	3	2	5.9E-04	-	0.55	3*1250.0
3	1	1	1.0E-04	-	0.15	1350.0
4	1	2	5.9E-04	-	0.55	1250.0
5	1	3	9.8E-04	-	0.30	1050.0
Total				0.83	0.00408	9500.0

For the sake of simplicity, the failure rate presented in Fig. 3.4 was used for all subgroups.

For this example problem a population of 30 chromosomes per generation where considered. The program was run for a total of 200 generations.

The program reports that the first chromosome is unfeasible because total cost exceeds the maximum cost ie. 9500.0 is greater than 8000.0. The chromosome is not selected for crossover. As it was mentioned before, the population of 30 chromosomes was run up to generation 200 ie. 6000 possibilities were scanned. The best result of the process was selected as the first solution. The computer print-out of the first solution is presented in Table 4.3 below.

Table 4.3. Computer print-out of 1st solution
Optimisation via a Genetic Algorithm

Program 'GAm' Version 1.5. Mar 2003
Example problem: e-Hub design

Options implemented:
Elitist Selection
Children replace parents

GA parameters:
Population size = 30
Chromosome length = 25
Max. number of generations = 200
Crossover probability = 0.1
Mutation probability = 0.01

Results:
*Fitness = 0.857
*Fitness probability = 7.59e-05
*Fitness cost = 6600
*Variables:
*n1 = 1
*λ1 = 0.0001
*n2 = 2
*λ12 = 0.0001
*n3 = 1
*λ13 = 0.0001
*n4 = 1
*λ14 = 0.0001
*n5 = 1
*λ15 = 0.0001

*Solution obtained in generation 24
*Total Number of Mutations = 1679
*Total Number of Crossovers = 748

The system failure rate for the first solution is 1.0 - 0.857 = 0.143 failures/100 years, ie. we can expect 14.3 failures per year.

The program was repeatedly run until the probability distribution of solution points was completed, ie. until $\sum_{\forall i} Pr_i$ reached 1.0. This happened at run number 4391. Fig. 4.6 shows the probability distribution of system fitness. Probabilities of identical fitness were added together to give the actual probability of that fitness happening. From the plot we can see that most values concentrate around three points: 0.848, 0.857 and 0.947. The minimum fitness is 0.848 the maximum is 0.947.

The Expected value of fitness is 0.911 with standard deviation $\sigma = 0.045$.

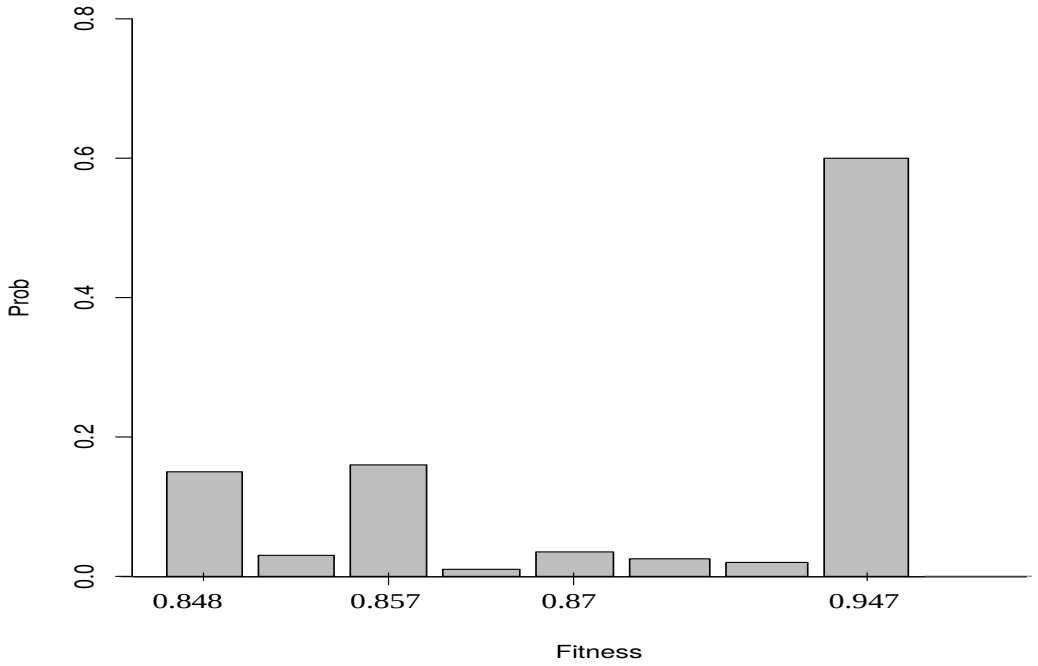

Figure 4.6: Probability Distribution of System Fitness.

The optimum solution from a reliability point of view is the configuration which produces the highest fitness (lowest system failure rate). The highest fitness is 0.947 which happens 2814 times (64%) out of the 4391 runs. Its associated probability is 0.598. Table 4.4 shows the optimum configuration.

Table 4.4. Optimum Configuration for the e-Hub

Subgroup	No. of elements	Element λ	Cost	Prob	Fitness	λ
1	2	0.00059				
2	1	0.0001				
3	1	0.0001				
4	1	0.0001				
5	1	0.0001				
Total			7900	0.598	0.947	0.053

Fig. 4.7 shows the configuration suggested by Table 4.4. The program allocates redundancies only to subgroup 1. The cheapest component is used in subgroup one, all other components are equally reliable. Total system failure rate is 1.0 - fitness = 0.053 (failures per 100 years). So for this configuration we can expect 5.3 failures per year. The total cost of this configuration is $7900 which is within the required range.

To see how the cost limit affects the result, the problem was run again for a required maximum cost of $10000.0. Fig. 4.8 shows the distribution of optimal results.

The best fitness of this distribution is 0.989 at a cost of $9750. This configuration happens 3 times out of 1303 runs (0.23%). The probability of this configuration is 0.00365. Notice that only 1303 runs were required to complete the distribution of optimal fitness because the higher cost constrain allows the program to scan more redundancies with lower and medium costs components, hence higher probabilities are considered in this case. Table 4.5 presents the configuration. This configuration was obtained in generation 137.

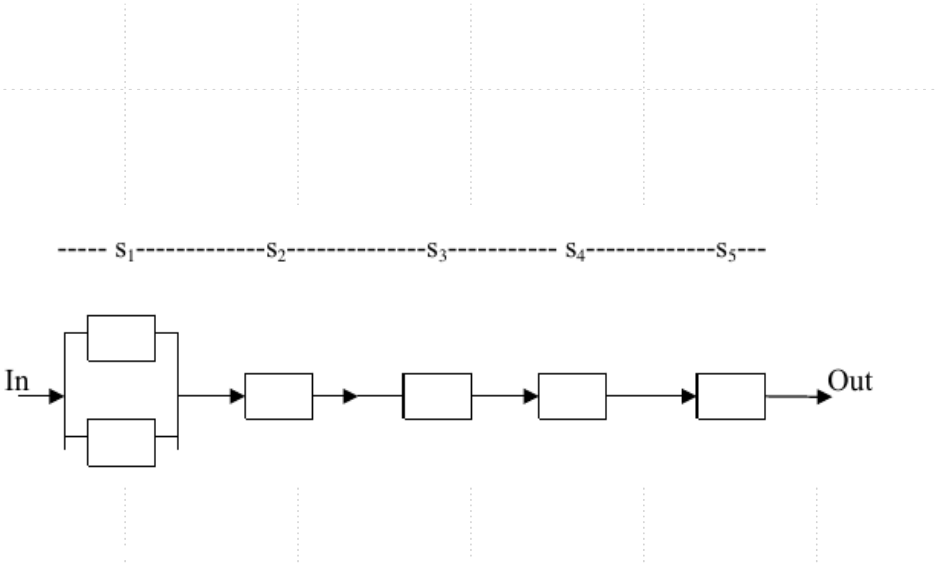

Figure 4.7: Optimum design for the e-Hub.

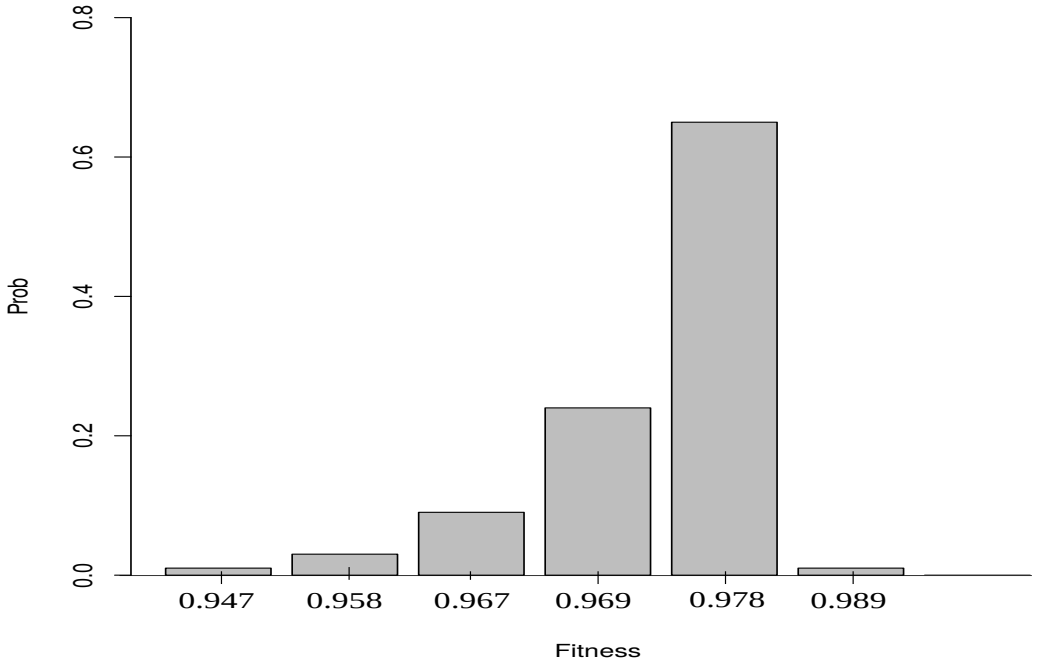

Figure 4.8: Distribution of optimal results for C_M = \$10000.0

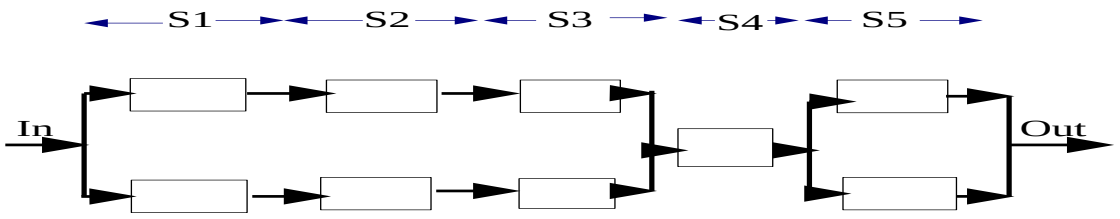

Figure 4.9: Optimal architecture for the e-Hub system at a cost of $9750.0

Table 4.5 shows that the GA improves system reliability by increasing the redundancies of subgroups 2, 3 and 5. The cost is increased by $1850, that is, by a 23.4% cost increase it is possible to reduce the failure rate from 5.3 to 1.1 failures per year.

Table 4.5. Optimum Configuration at a cost of $9750.0

Subgroup	No. of elements	Element λ	Cost	Prob	Fitness	λ
1	2	0.00098				
2	2	0.00098				
3	2	0.00098				
4	1	0.0001				
5	2	0.00098				
Total			9750	0.00365	0.989	0.011

The optimal architecture for the e-Hub system described in Table 4.5 is shown in Fig. 4.9.

In the problem used to illustrate this chapter only optimisation of redundancies and component reliability are considered. It is however possible to also consider optimisation of maintenance intervals for an e-Hub server. In some problems the maintenance cost must also be considered so an optimum solution will also include maintenance over the lifetime of the server.

4.6 References

Jasmon G.B. and Kai O.S. (1985). A New Technique in Minimal Path and Cutset Evaluation. IEEE Trans on Reliability. Vol. R-34, No. 2. 1985 June.

Goldberg D.E. (1989). Genetic Algorithms in Search, Optimisation, and Machine Learning. Addison-Wesley Publishing Co.

Gen M. and Runwei C. (1997). Genetic Algorithms and Engineering Design. A Wiley-Interscience Publication.

Pham D.T. and Karaboga D. (2000). Intelligent Optimisation Techniques. Springer London.

Coit D.W. and Smith A.E. (1996). Reliability of Series-Parallel Systems Using a Genetic Algorithm. IEEE Trans. on Reliability. Vol. 45, No. 2. June 1996.

Painton L and Campbell J. (1995). Genetic Algorithms in Optimisation of System Reliability. IEEE Trans. on Reliability. Vol 44, No. 2. June 1995.

Soh B. and Sattar S. (2002). Performance and Reliability Study of an e-Hub in Collaborative B2B e-Commerce. Dept. of Computer Science and Computer Eng. La Trobe University.

Sanabria L.A., Soh B., Dillon T.S. (2002). Reliability Analysis of an Intranet Server Using the Cut-Set Technique. Computer Science and Computer Eng. Department. La Trobe University.

Sharma J. (1976). Algorithm for Reliability Evaluation of a Reducible Network. IEEE Trans. on Reliabililty. Vol. R-25, No. 5. Dec. 1976.

Chapter 5

Modelling natural phenomena

In this chapter models to study natural phenomena will be discussed. The aim of these types of models is to assess the risk posed to people and the build environment by sudden-impact natural phenomena like wind, precipitation or temperature. Combinations of these phenomena can result on floods or wild fires. In this chapter we focus on wind, wild fires will be discussed in the next chapter.

A fundamental characteristics of these natural phenomena is that they are impacted by climate change. To consider the impact of climate change it is necessary to use data produced by climate models. This issue will be presented in Chapter 7.

In the Australian context wind is probably the most dangerous natural phenomena. Severe winds are responsible for about 40% of damage to Australian residential buildings. The Impact of wind on Australian houses is significantly higher than for other natural hazards such as floods (22%), wild fires (19%), and earthquakes (6%) (Chen, 2004).

5.1 Introduction

Unlike the models discussed in the previous chapters, models to study natural phenomena are based on data, these types of models are referred to as Statistical Models. initially we will study models based on wind observations, that is, records of wind collected over many years by Meteorological organisations. The models will be extended later to consider the impact of climate change by using data produced by climate models, these types of data are termed 'climate simulations'.

One of the main applications of wind hazard models is in the production of regulations for constructions of buildings. In Australian these regulations are compiled in the Australian and New Zealand standard for structural design actions AS/NZS 1170.2:2011 (2011). These regulations set out precedures for the design of structures subjected to wind loads. For this, it is important to calculate wind speeds as accurately as possible (Ginger et al. 2013)

Since the values of natural phenomena are recurrent the quantification of the hazard from natural phenomena is achieved by calculating the frequency with which extreme values repeat and the maximum they can reach. The main indicator of natural hazard is called the Average Recurrence Interval (ARI) more commonly known as Return Period (PR). AS/NZS 1170.2:2011, for instance, prescribes that structures in the AS/NZ region must be designed to stand wind loads corresponding to a RP of 500 years. The Argentinian standards require that buildings type III be designed for wind loads corresponding to a RP of 1300 years (Natalini 2016).

Natural phenomena are not stationary, they change over time so wind and other standards for building construction must be updated from time to time. That is why is so important to consider the impact of climate change in order to set up regulations for realistic levels of hazard.

Results from these types of models are also important for emergency authorities who must develop contingency plans to deal with disasters produced by natural phenomena. For them it is important to allocate scarce resources where they are needed the most and build up preparedness in the regions exposed to high levels of hazard. Planning authorities are also interested in natural hazard studies, they must plan expansion of cities to areas not likely affected by natural disasters.

5.2 Brief introduction to the Statistics of Extreme Values

The aim of an extreme value (EV) analysis is the estimation of the probability of events that are more extreme than any that have already been observed. To illustrate this definition suppose that we want to build an observation tower in a given airport for a working life span of 50 years. Construction work is planned to start in the year 2020. Suppose that we have only 20 years of wind speed records in this airport. We need to design the structure to stand the maximum wind load which may be found in the 50 year span. The problem is, then, what is this value? EV anaysis provides a framework to extrapolate the given data to values well beyond the available records and hence helps to answer the question.

In mathematical terms the problem can be expressed as, find M_n, the maximum of a process measured on a regular time scale. In our case the process is wind speed.

$$M_n = max[X_1, X_2 \cdots X_n] \tag{5.1}$$

Where $X_1, \cdots \mathrm{X}_n$ is a sequence of independent random variables having a common distribution F. M_n represents the maximum of the process over 'n' time units of observations. If 'n' is the number of observations in a day, then M_n corresponds to the daily maximum value. To calculate F we can look at the asymptotic behaviour of F^n as n $\to \infty$. For this, M_n must be normalized (similarly to what we did in eq (2.19)) using the expression,

$$M_n^* = \frac{M_n - b_n}{a_n} \tag{5.2}$$

Where a_n and b_n are suitable constants which stabilize the location and scale of M_n as 'n' increases. Coles (2001) shows that the limiting distribution of M_n^* is the family of Extreme Value distributions G(z), given by,

$$I : G(z) = exp(-exp[-\frac{z-b}{a}]), \qquad -\infty < z < \infty \tag{5.3}$$

$$II : G(z) = \begin{cases} 0, & z \leq b \\ exp(-[\frac{z-b}{a}]^{-\alpha}), & z > b \end{cases} \tag{5.4}$$

$$III : G(z) = \begin{cases} exp(-[-(\frac{z-b}{a})^{\alpha}]), & z < b \\ 1, & z \geq b \end{cases} \tag{5.5}$$

for parameters $a > 0$. b and $\alpha > 0$ for families II and III.

These families of EV distributions are known by the names, I = Gumbel, II = Fréchet, and III = Weibull distributions.

The theory behind EV statistics is similar to the Central Limit Theorem (CLT); both infer the limiting distribution of independent, identically distributed (iid) random variables. According to the CLT, the mean value of a sample of iid random variables converges to a standard normal distribution. Similarly if the maxima of a large number of iid random variables converge to a distribution, this distribution has to be a member of the Extreme Value Distributions (Jagger and Elsner, 2006).

Fig. 5.1 illustrates the case. In Fig. 5.1A the histogram of mean values of 2000 datasets, each containing 1000 samples generated at random between -1 and 1, is shown. A normal (black line) and a Generalised Extreme Value distribution type III (blue line) are fitted to the histogram. It is clear that the normal distribution is better at fitting the histogram of mean values than the Generalised Extreme Value distribution type III. Figure 5.1B shows the histogram of maxima, this time the GEV is better at fitting the histogram of peak values than the normal distribution.

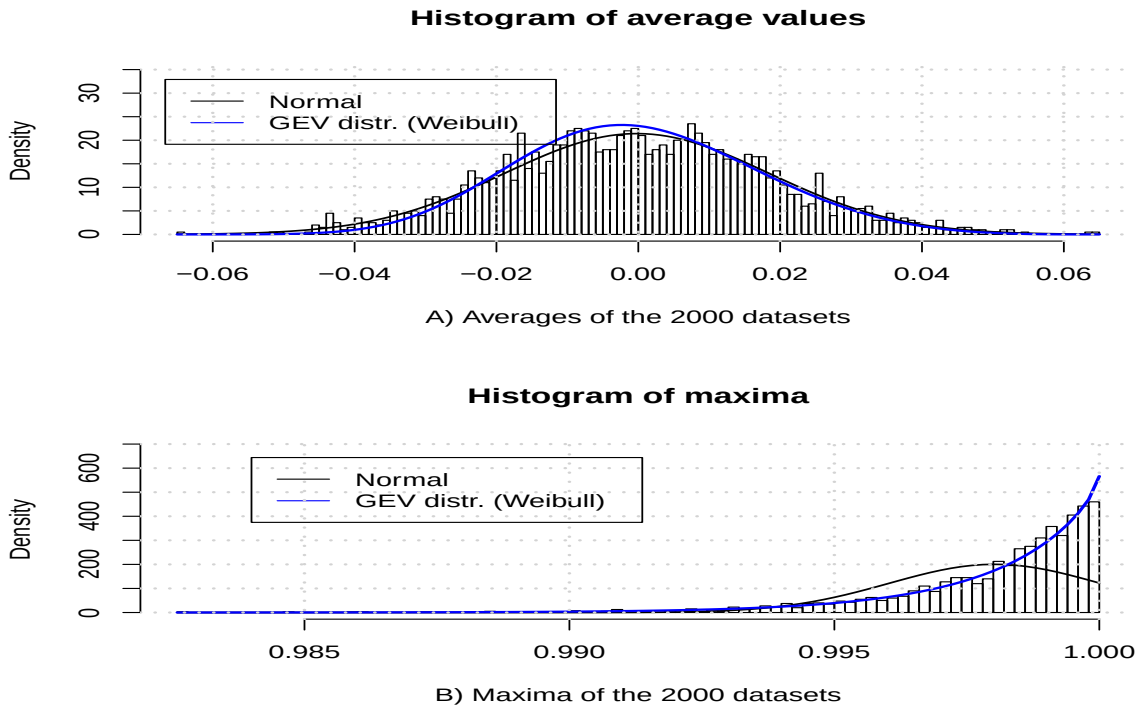

Figure 5.1: Histogram of a) mean and b) maxima of 2000 datasets

The three different forms of the family of EV distributios were combined into a single expression by Jenkinson (1955) which greatly simplifies the analysis. The expression, known as the Generalized Extreme Value distribution (GEV), is given by,

$$G(z) = exp(-[1 + \xi * (\frac{z-\mu}{\sigma})]^{1-\xi}) \tag{5.6}$$

defined on the set {z: $1 + \xi * \frac{(z-\mu)}{\sigma} > 0$}, where the parameters satisfy $-\infty < \mu < \infty$, $\sigma > 0$ *and* $-\infty < \xi < \infty$.

5.3 Return Period

To illustrate the calculation of the RP we will be using datasets of wind speeds acquired from the Australian Bureau of Meteorology (BoM 2020), from the stations located in the Sydney region. Fig. 5.2 shows the histogram of daily maximum wind speed at Sydney Airport. For this study we have selected 30 years from 1990 since the latter part of the records are usualy more reliable (Cechet & Sanabria 2011). There are 10776 values in the data selected with a maximum of 33.4 m/s. It is particularly important to note that the histrogram is not symmetric, the figure shows a long right hand side tail usually called a 'hard tail'. This is characteristic of all natural phenomena, the figure is skewed by the presence of extreme values. In the case of drought the hard tail will be located to the left.

To calculate the RP consider again eq (5.1). The data $X_1, X_2 \cdots$ is blocked into sequences of observations of length 'n'. For some large value of 'n' a series of block maxima are generated $X_{n,1}, X_{n,2} \cdots X_{n,m}$. The usual time period for these series of block maxima is one year, in this case we are dealing with annual maxima, i.e. the maximum value of the block of one year of observations. The GEV is then, fitted to the series of maximum values. This methodology is called "block maxima" as will be discussed in Section 5.4.

A quantile z_p corresponding to the GEV distribution of eq (5.6) is given by inversion as,

$$z_p = \begin{cases} \mu - \frac{\sigma}{\xi} * [1 - (-log(1-p))^{-\xi}] & for \quad \xi \neq 0, \\ \mu - \sigma * log(-log(1-p)), & for \quad \xi = 0 \end{cases} \tag{5.7}$$

Where $G(z_p) = 1$ - p.

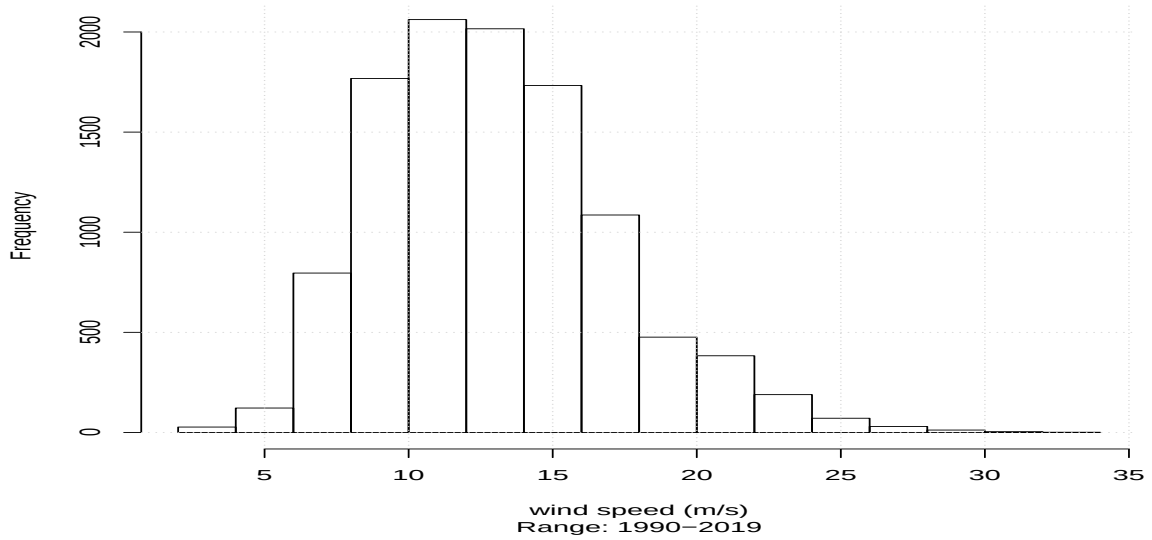

Figure 5.2: Histogram of wind speed at Sydney Airport

z_p is the return level associated with the return period 1/p, i.e. z_p is expected to be exceeded on average once every 1/p years (Coles, 2001). In statistical terms, the probability of exceeding z_p in any particular year is 'p'. This probability can be calculated using the exceedance distribution, Exc(x), introduced in Section 1.4. Fig. 5.3 presents both the exceedance and the distribution functions, the arrow shows the probability of exceeding the value 60. The return level corresponding to a RP of 'yr' years can be calculated as,

$$RP(yr) = \frac{1}{Exc(yr)} \tag{5.8}$$

The series of annual maxima from Fig. 5.2 was determined and the corresponding values for the RP were calculated using eq. (5.8) as shown in Fig. 5.4. Note that the x-axis, usually years, is logarithmic. It shows that the RP corresponding to a wind speed of 32 m/s is 10 years, i.e. a wind of 32 m/s in the Sydney Airport is exceeded, on average, once every 10 years. The emphasis here is on the expression *on average*, we are not saying that a wind speed of 32 m/s is exceeded *exactly* every 10 years.

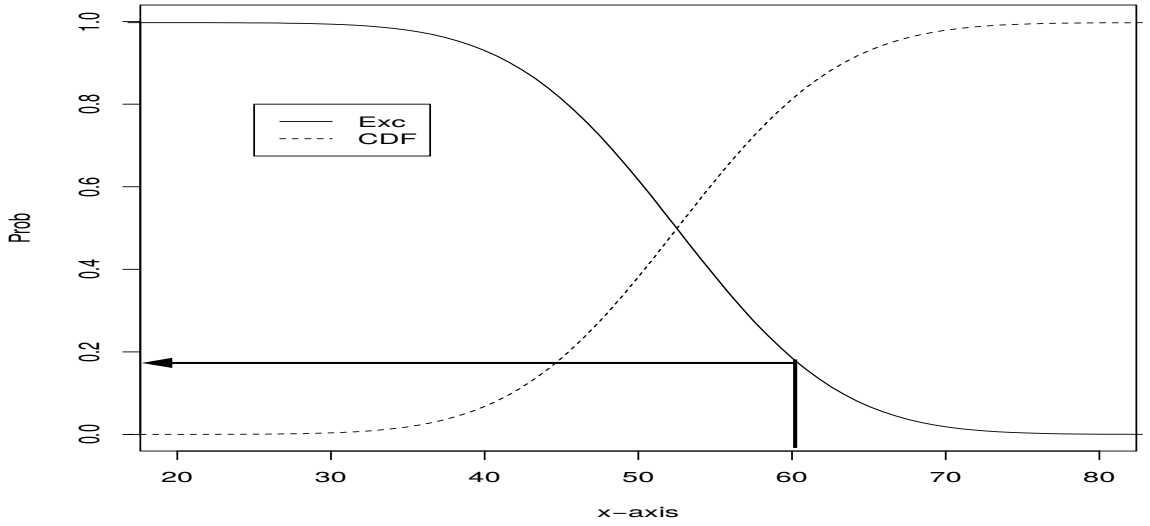

Figure 5.3: Cumulative and Exceedance distributions

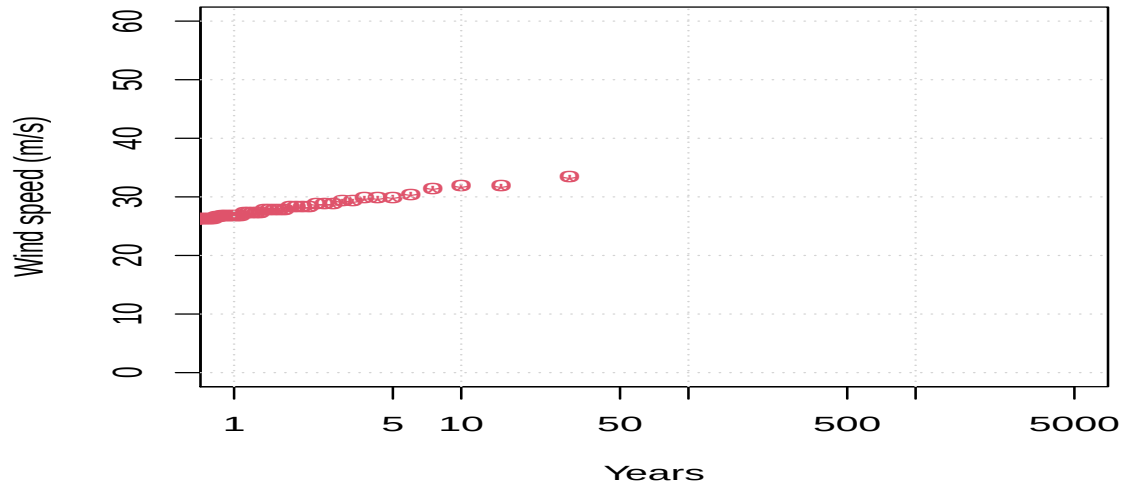

Figure 5.4: RP of observed wind speed (Sydney Airport)

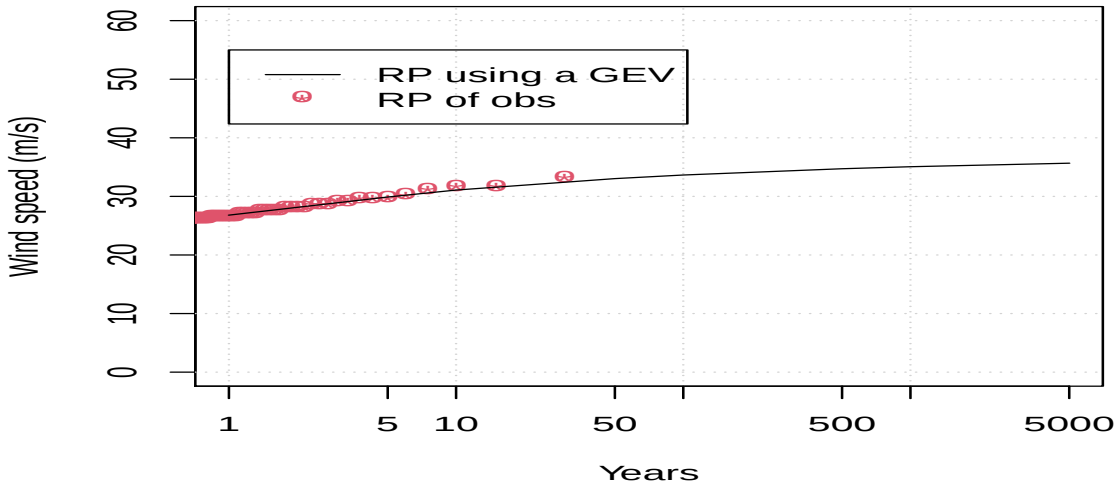

Figure 5.5: Curve of RP of wind speed using the GEV

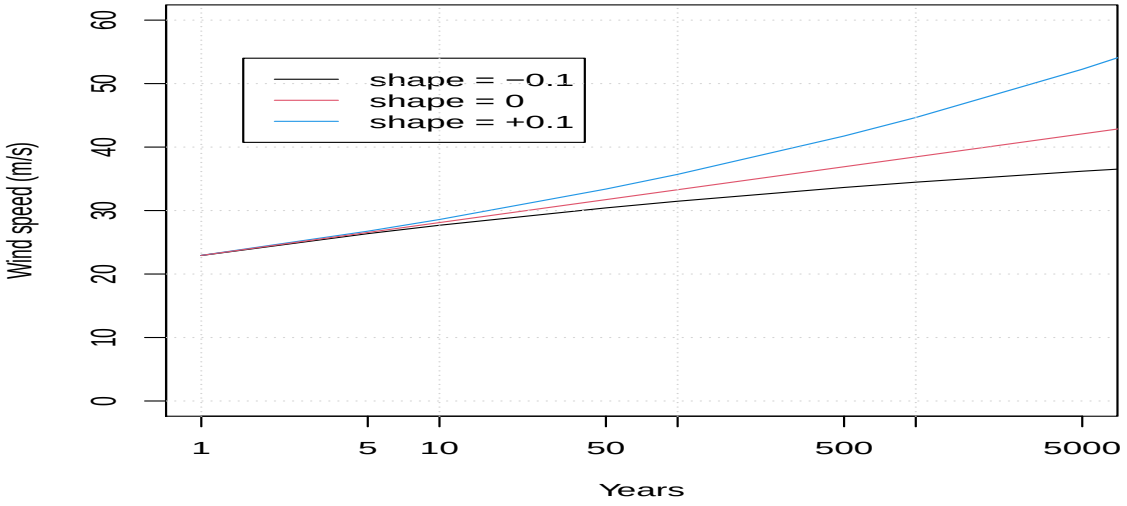

Figure 5.6: Sensitivity of GEV models to variation of the shape parameter

The curve of Fig. 5.4 shows the RP of *observed* wind speeds, that is why it covers a range of only 30 years. For this reason these types of curves are almost useless, they do not give new information. Remember, the Australian/NZ standards prescribes the wind load for design of structures in the region as the speeds corresponding to a RP of 500 years. What we need is, then, a model which allows us to project the curve to a range of years beyond the available data, this is the role of the extreme value distributions.

To fit the GEV distribution to (the tail of) the observations shown in Fig. 5.2 it is necessary to calculate 3 parameters: Location (μ), scale (σ) and shape (ξ) as presented in eq. (5.6). There are several methods to fit the GEV, the most effective ones are the method of moments and the maximum likelihood method (MLE) (Oztekin 2005; Prescott and Walden 1980; Seguro and Lambert 2000). Fig. 5.5 presents the curve of RP calculated by fitting the GEV by MLE (Stephenson 2004), this curve has been superimposed over the curve of observed data (the red dots). With this model we can calculate RP of 5000 years and beyond. Note how well the GEV fits the data. The accuracy of these types of calculations will be discussed later.

To finish off this part let us observe the role of the shape parameter ξ in extreme value distributions. The sign of this parameter defines the shape of the curve. It is convex for $\xi < 0$, it is concave for $\xi > 0$. For $\xi = 0$ the curve is linear. Fig. 5.6 illustrates this characteristic of GEV models. This is a very important observation because positive or zero shape parameters make the function unbounded. For this reason some researchers recommend the use of GEV type III (Weibull) distributions ($\xi < 0$) for modelling natural phenomena (Lechner et al. 1992; Holmes 2007). These types of phenomena are naturally bounded by laws of physics, you can have very high levels of precipitation or wind speed but they are always bounded by an asymptotic value. This value can be calculated from the available data, using the expression (Coles 2001),

$$asympt = \mu - \sigma/\xi \qquad for \quad \xi < 0 \tag{5.9}$$

These days fitting distributions and other type of modelling work is greatly facilitated by the availability of computer software. Our favourite software for statistical analysis and modelling is the R package (R Core Team 2018).

5.4 Fitting EV using the 'Peaks-over-threshold' technique

The discussion of RP curves so far considers only annual maxima as explained in Section 5.3. Most Meteorological Offices record wind speed, and other climatic variables, at different intervals. So you can obtain datasets of 1-minute, 3-hourly, hourly and maximum daily wind speed, temperature or other variables. In some applications the interval of importance is annual maxima, for instance if we want to study the hazard produced by sea level raise or riverine floods (Twan 1992; Kjeldsen 2014).

For wind hazard is more common to use maximum daily. In this case the techniques presented above are not appropriate. One reason is that a lot of data is wasted. Fitting the EV distribution to series of only one value per year may introduce inaccuracies because you can have several higher values in a particular year than the maximum of another year. On the other hand the annual maximum of two consecutive years may not be independent from each other, one of the conditions for extreme value analysis. For these reasons two different ways to fit data to extreme value distributions have been developed: "block maxima" and "peaks-over-threshod' (Palutikof 1999).

In the "block maxima" technique, the data is split into groups of annual values, then the maximum value of each block is calculated and a GEV is fitted to these series of maxima as was explained in Section 5.3.

The "peaks-over-threshold" technique utilizes all values over a given threshold, in this case the extreme value distribution used is the Generalized Pareto Distribution (GPD). This methodology has a number of advantages over the "block maxima" method; firstly it uses a lot more data to fit the distribution, and secondly, by setting the threshold high enough, the data will be better distributed in time, improving the chances that the data samples are independent from each other, one of the conditions of EV applications, as explained before. This is the method recommended when maximum daily observations are available (Coles, 2001; Holmes and Moriarty, 1999)

The GPD is defined by the expression,

$$H(y) = 1 - (1 + \frac{y * \xi}{\hat{\sigma}})^{-1/\xi} \tag{5.10}$$

defined on the set $\{y : y > 0, \quad \text{and} \quad (1 + y * \xi/\hat{\sigma}) > 0\ \}$,
where $\hat{\sigma} = \sigma + \xi * (u - \mu)$ and u is the threshold.

Note the similarity of expressions (5.6) and (5.10). If a given "block maxima" have a limiting distribution "G(z)", then threshold excesses have a corresponding distribution "H(y)". More importantly, the parameter ξ is the same for both distributions. The shape and the other parameters of the GEV are very sensitive to the block size "n"; however in GPD calculations ξ is invariant to block size, while σ is insensitive to changes in μ and σ (Coles 2001), hence all figures of Section 5.3 are also valid for this section.

For calculation of the RP using the GPD, let us suppose that expression (5.10) has been fitted to some data. Then, for $x > u$,

$$Pr[X > x|X > u] = [1 + \xi * \frac{(x-u)}{\sigma}]^{-1/\xi} \tag{5.11}$$

It follows that,

$$Pr[X > x] = \zeta_u * [1 + \xi * \frac{(x-u)}{\sigma}]^{-1/\xi} \tag{5.12}$$

Where, $\zeta_u = Pr[X > u]$

Hence the level x_m that is exceeded on average once every "m" observations is the solution of,

$$\zeta_u * [1 + \xi * \frac{(x_m-u)}{\sigma}]^{-1/\xi} = 1/m \tag{5.13}$$

That is,

$$x_m = u + \frac{\sigma}{\xi}[(m * \zeta_u)^{\xi} - 1], \tag{5.14}$$

provided "m" is sufficiently large to ensure that $x_m > u$; and $\xi \neq 0$. For the case in which $\xi = 0$, the corresponding expression is,

$$x_m = u + \sigma * log(m * \zeta_u) \tag{5.15}$$

One of the problems found in fitting a GPD to given data samples is the selection of the appropriate threshold value u. High threshold values result in the selection of only a few data points, most likely not enough for a good fitting of the distribution. Low values result in too many samples which are most likely not independent from each other. On the other hand return period calculation using GPD distributions are very sensitive to the threshold selection as presented in Fig. 5.7.

Although there are methods to help modellers select the appropriate threshold for a given dataset they are mostly visual, subjective techniques, prone to producing inaccurate results and inappropriate for large scale applications as in gridded data. To model wind speeds using GPD distributions is necessary to develop a technique for automatic selection of the appropriate threshold for a given dataset. This is the topic of the next section.

5.4.1 Automatic selection of the appropriate threshold

The algorithm was developed by observing that there are two different sub-sets of curves in Fig. 5.7: Sub-set 1 (with u = 20, 22.5, 23, 23.5) m/s is characterised by a shape parameter greater or equal to 0 which produces unbounded curves. As we explained before these types of curves are inappropriate for modelling naturally bounded phenomena.

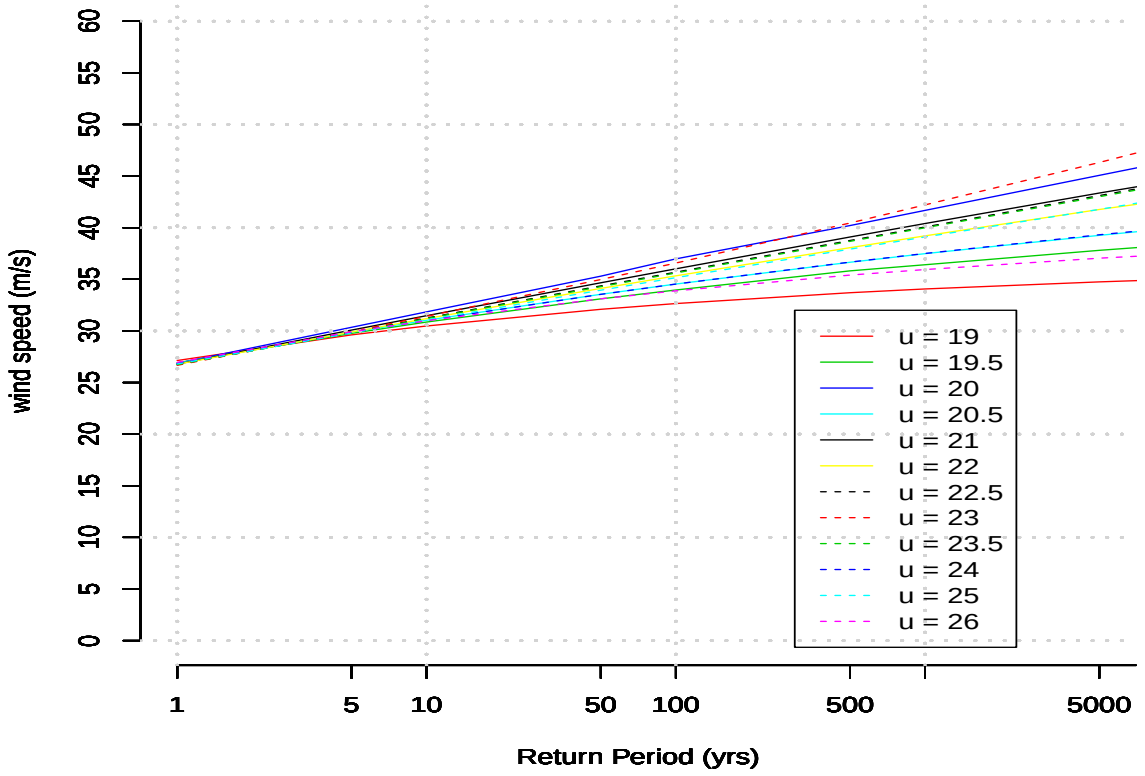

Figure 5.7: Sensitivity of the GPD curve to threshold selection

Sub-set 2 (u = 19, 19.5, 20.5, 21, 22, 24, 25, 26, 27) m/s is characterised by a negative shape parameter which results in bounded curves appropriate for modelling wind speed; for this reason we call this the sub-set of feasible curves. Notice however that the curves produced by thresholds 19, 25 and 26, are flat with quick convergence to a very low speed value and hence are not appropriate for modelling wind hazard which depends on the high wind speed values. From Fig. 5.7 it is clear that the most appropriate threshold from those shown is u = 21 (black curve) which produces a bounded curve with the highest return period.

The algorithm generates a sub-set of *feasible* RP curves in steps of 0.5 (m/s) starting with a value close to one-third of maximum wind speed. It returns the appropriate threshold for modelling the given dataset within this sub-set. It is generally the threshold producing the highest return period curve. Based on return periods of observed wind speeds generated for a number of BoM datasets, a set of rules for selection of the appropriate threshold were compiled and coded to produce the automatic algorithm (Sanabria & Cechet, 2007). The algorithm was tested with the same dataset presented in Fig. 5.2. As expected the algorithm returns 21.0 (m/s) as the appropriate threshold to fit the GPD to the given dataset as it produces the bounded RP curve with the highest asymptote (the black curve in Fig. 5.7).

5.4.2 Confidence Interval

We have already mentioned the need for assessing the accuracy of the RP curves presented above. Statisticians have developed a technique to do this, it is called a Confidence Interval. A confidence interval (CI) shows the range of values in which the true value of the RP level lies for a given probability. In this work we are interested in finding confidence intervals with 95% probability. That is, intervals with a 95% probability that the true return period level lies within the interval.

The confidence interval depends on the size and structure of the dataset, particularly the variance-covariance matrix, which measures the spread of the samples around their mean. There are two basic algorithms for calculation of the confidence intervals of curves produced by extreme value distributions: The Delta method and the Profile Likelihood method. Both methods have been implemented in the R environment by Gillelland and Katz (2009) based on Coles (2001). Applying the methods to temperature data, they found out that the Profile-likelihood method gives better results because it considers the asymmetry of the data (Gillelland and Katz, 2005).

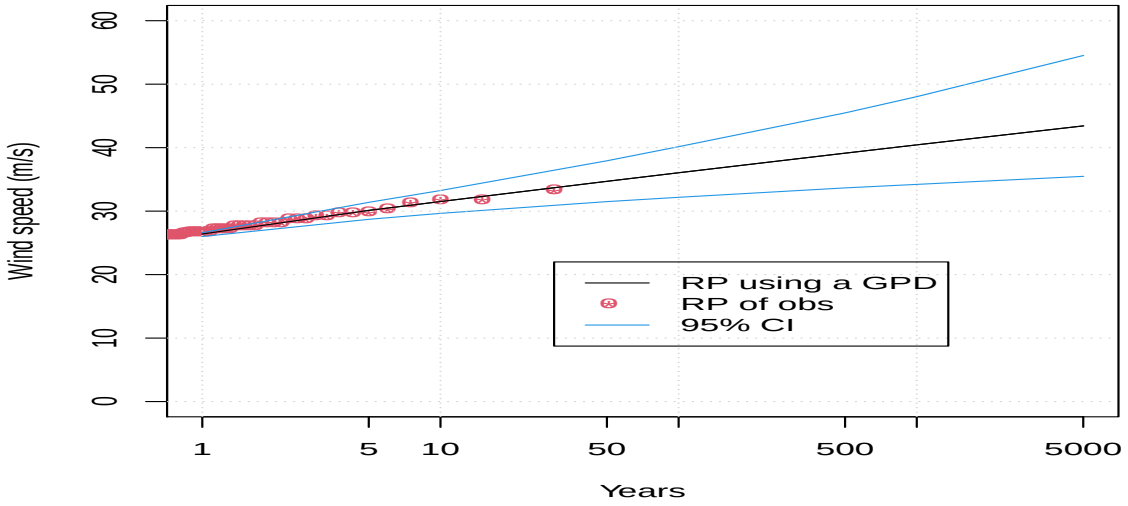

Figure 5.8: RP of wind speed with 95% Confidence Interval

As an illustration Fig. 5.8 shows the RP curve for the Sydney Airport wind speed with the 95% confidence interval calculated using the Profile-likelihood method (since wind speed is highly asymmetric). The red points are the RP of the observed wind speeds. Since our RP curve lies inside the CI we can say that there is a 95% probability that our calculations are correct.

Notice that the confidence interval increases at high RP values indicating a higher degree of uncertainty when making inferences far beyond the range of the data (30 years for Sydney Airport).

We had already mentioned that the GPD produces more accurate results because it uses a lot more data to fit the extreme value distribution (Holmes and Moriarty 1999). This can be observed by comparing figures 5.5 and 5.8 (black curve). The 1000 years RP for the former is about 35 m/s while the corresponding value for the later is 40.4 m/s an increase of more than 12%. Note also that the GPD curve fitts better the red points.

5.4.3 Diagnostic plots

The R package "evd" (Stephenson 2004) allows the analyst to visually assess the quality of the GPD fitting by comparing against a series of standard plots as shown in Fig. 5.9.

The top left curve is the so called QQplot, the curve of quantiles modelled using the GPD (x-axis) against the actual quantiles (calculated from the observations; shown in the y-axis). A quantile is the value of the x-axis corresponding to a given probability (y-axis) of the cumulative distribution function (CDF). The QQplot allows the analyst to quickly assess the quality of the fitting: if the circles lie on the diagonal line the fitting is perfect. Note that in this case most circles lie on the diagonal line except for the extreme values (ws $>=$ 32 m/s). The top right hand curve is basically the same QQplot but this time only quantiles greater than the threshold (21 m/s) are considered. This plot also gives you the 95% CI and the curve of linear regression between the modelled quantiles and the actual ones (the light blue line). The bottom left curve shows the probability density function (pdf) of the wind (black curve) and its corresponding fitting using the GPD (blue dotted curve). Notice that the fitting of the extremes is very good while the fitting of the main body of the distribution is not so good; as explained before, this is the central characteristic of extreme value distributions. The bottom right curve is the curve presented in Fig. 5.8. From these plots we can conclude that the fitting of the extreme values of Sydney Airport wind speed using the the GPD is very good and hence our selection of the threshold was correct.

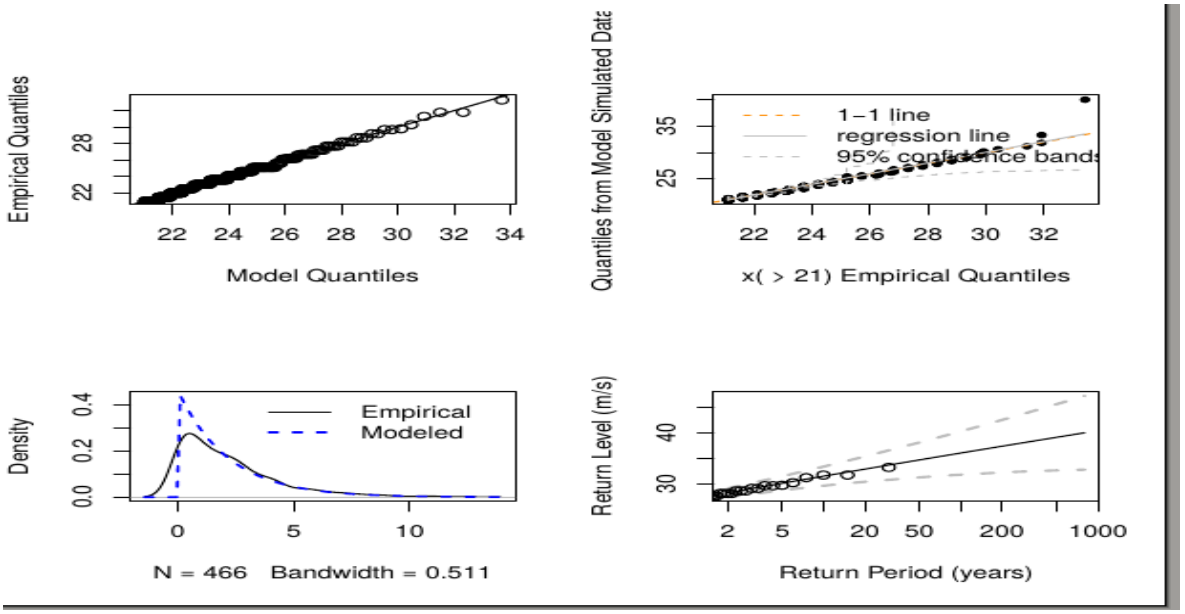

Figure 5.9: Diagnostic plots

Another problem found in fitting extreme value distributions is that the data samples need to be independent from each other otherwise the technique will produce erroneous results. This could be a serious limitation for precipitation or wild fire studies because these phenomena are associated with weather systems that may persist for a few days (i.e. the same event with high values may span several days at more than one weather station location). Another example is the case where wind speeds from different stations are lumped together to produce what Holmes calls a "super-station"(Holmes 2007). This technique is useful when there are stations which have been in operation for only a few years and have not produced enough data for extreme value analysis. Lumping together several stations can also be necessary to calculate the wind hazard over a *region* comprised by the stations rather than at a specific location. The dataset made up by joining several datasets must comprise independent events if we want to fit a GPD to it. An efficient technique to scan a dataset for independent events is presented next.

5.4.4 Clustering wind records

Datasets of meteorological observations over a region that have similar wind speed distributions can be joined into a single dataset (Holmes 2007). For these cases it it necessary to develop an algorithm that locates those observations of the joined dataset that were caused by the same phenomenon. To guarantee independence they must be counted only once and duplicate samples should be deleted. In this section we have used the method developed to cluster precipitation observations into independent events by White et al (2010). A brief description of the method follows: First the joined dataset is sorted in chronological order and then the dataset is grouped into clusters by defining a minimum threshold u. Consecutive exceedances of u are assumed to belong to the same cluster. A cluster is deemed to have terminated when r values fall below u. The next value over u starts a new cluster which terminates when r observations fall under u and so on. The procedure is illustrated in Fig. 5.10 top. In this figure u = 10 and r = 4. The first cluster starts at sample 10 and terminates at sample 12 because it is followed by 7 values < 10. The next cluster starts at sample 19 and terminates at sample 22 which has a value of 5.1, and so on. The new time series of independent events is the maximum value in each cluster, as shown in Fig. 5.10 bottom: the set {C1, C2, C3, $\cdots$ } To illustrate the methodology consider 3 weather stations located in southern

NSW (Australia). These stations have been selected because they have similar wind characteristics and are closed to each other (Sanabria and Cechet 2015). The stations are: Williamstown (WILL), Bankstown (BANKS) and Richmond (RICH). the datasets of these stations were joined into a single dataset which has 36313 samples over 139 total years: [1968-1992], [1942-1994] and [1942-2005] respectively. Next the joined dataset has to be examined for non-independent events. Table 5.3 presents a small section of the ordered joined dataset. The Columns show the date, the wind speed and the station name (where the observation was recorded) respectively.

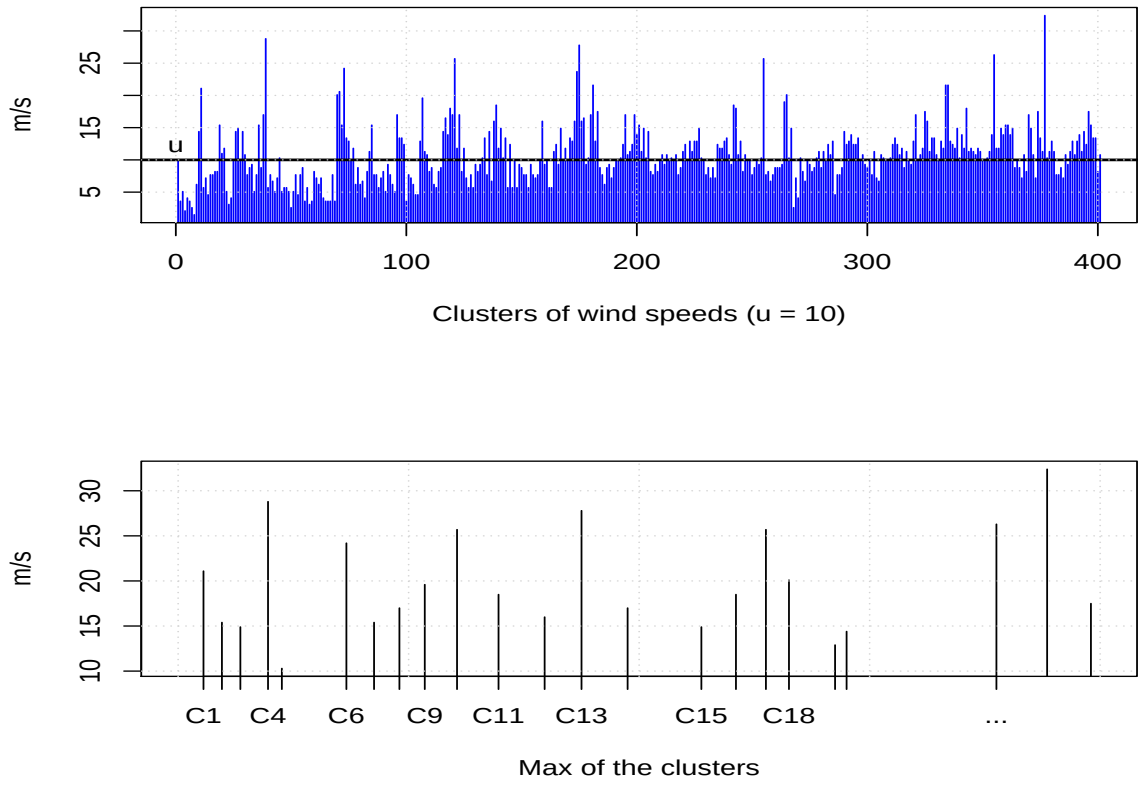

Figure 5.10: Clustering wind data into independent samples

Table 5.1. A small section of the joined dataset in chronological order.

Date	Wind speed	Station
1978-03-19	26.3	BANKS
1978-03-19	25.2	WILL
1978-06-01	30.9	WILL
1978-06-02	28.3	BANKS
1978-06-02	26.3	WILL
1978-07-06	25.7	WILL
1978-12-12	27.3	WILL
1979-01-07	25.2	BANKS
1979-11-15	37.1	BANKS
1979-11-23	26.8	BANKS
1979-11-26	31.4	BANKS
1980-06-29	27.8	WILL
1980-08-31	25.7	BANKS
1980-08-31	25.2	RICH
1980-08-31	28.3	WILL
1980-09-15	26.3	BANKS
1980-09-15	26.3	RICH
1980-09-15	27.8	WILL
1980-10-12	26.8	WILL
1981-01-21	25.7	RICH
1981-07-29	25.7	WILL

Consider the first 2 rows: they show that on 1978-03-19 BANKS recorded 26.6 m/s while WILL recorded 25.2 m/s; it is very likely that both records were produced by the same meteorological phenomenon. Further down in row 13 BANKS recorded 25.7 on 1980-08-31. On the same day RICH recorded 25.2 and WILL 28.3, again it is very likely that those records were produced by the same meteorological phenomenon and hence cannot be considered independent from each other. Similarly BANKS recorded 26.3, RICH 26.3 and WILL 27.8 on 1980-09-15, all of similar magnitude and likely to be produced by the same phenomenon.

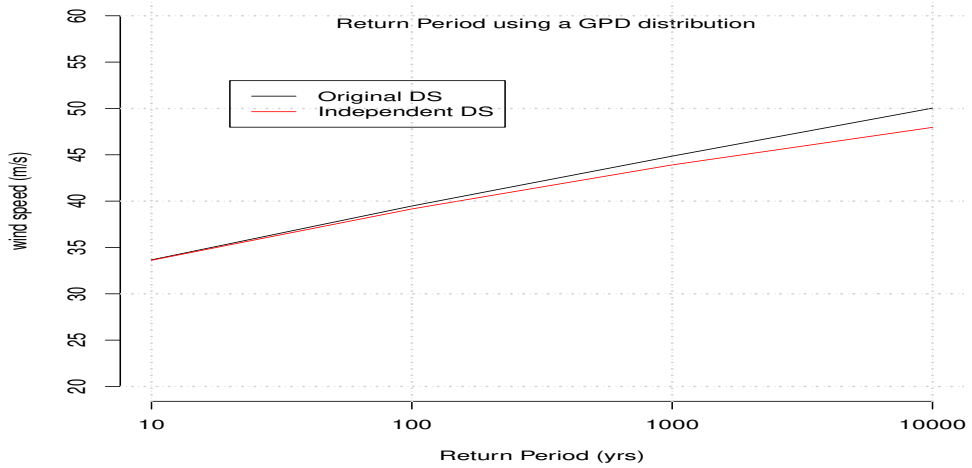

Figure 5.11: RP of clustered datasets: Original and independent

The algorithm clusters contiguous samples in to one single event which becomes a sample of the new independent dataset, see Fig. 5.10 bottom. In our case we defined a value of u = 20 m/s as the minimum threshold to select samples for the new time series; i.e. only samples > 20 are considered. A gap between clusters r = 4 was used, i.e. a cluster is deemed to have finished when 4 observations fall below 20. Note that r is given in days since we are using maximum daily gust wind speeds. The maximum value in the cluster is selected for the new time series. There are 1552 samples exceeding 20; a total of 899 clusters of independent samples were found, that is, only 899 of the 1552 samples exceeding 20 are deemed to be independent from each other. Now we can fit a GPD to this dataset in order to calculate the regional RP.

Fig. 5.11 compares the RP produced by the grouping algorithm (the red line) with the RP from the original dataset (DS) i.e. The original grouped dataset without clustering into independent events (black line). The latter joins the individual datasets without investigating whether they are independent from each other. Fig. 5.11 shows that the methodology discussed here produces results lower than the results of the non-independent dataset, particularly for high return periods. This is because the algorithm to select independent events eliminates high values of wind speed found in different stations that are likely to be produced by the same phenomenon. An accurate calculation of RP has the potential to save millions of dollars to the building construction industry.

5.5 Risk analysis

The discussion presented so far includes only hazard analysis, i.e. we have only considered the danger posed by wind speed, for risk analysis is it necessary to also consider the actual damage produced on people and the build environment by the hazard. Risk is a non-linear function of hazard, exposure and vulnerability, each of these elements are characterized by a probabilistic function and hence risk is the result of the convolution of these 3 elements. As explained in Section 1.3 solution of these types of operations is complicated hence most researchers solve the problem by numerical approximation. The methodology developed in Geoscience Australia (GA) for assessment of risk will be presented next.

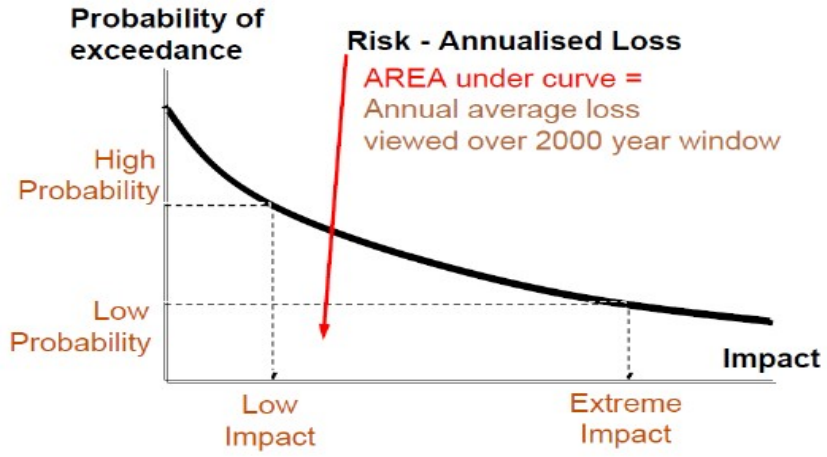

Figure 5.12: Probable Maximum Loss curve (PML)

GA's methodology for risk and impact analyses is a numerical convolution of Hazard, Exposure and Vulnerability. Each element of this paradigm is required to ensure comprehensive and integrated assessment within a consistent framework (Cechet et al. 2012).

Exposure refers to elements at risk from a hazard including residential buildings, people, infrastructure, industries, crops, or natural habitats such as forests or reef systems. Realistic studies of severe wind risk should also include other infrastructure and assets such as commercial and industrial buildings, agricultural crops, and power and telecommunication networks to ensure a complete understanding of the risk from severe winds.

Exposure, when referring to elements in the built environment, can be defined in a building-specific format, where each individual structure in the community is identified and characteristics such as geographic location, age, wall type and roof type are known. Alternatively, a statistical definition can be used, where the numbers of buildings in a specified area is known, and their distribution amongst a range of types is also known. In the former situation, we require high-resolution information on the level of hazard at the location of each building. The latter definition allows the use of more generalised information, which is representative of the sampled area.

Vulnerability describes the capacity of exposed elements (buildings, infrastructure or people) to withstand, and recover from, the impact of hazards. The vulnerability relates gust wind speed to the damage ratio, the repair cost divided by the replacement cost, for the particular type of building in question. The closer the damage index is to one, the more heavily the building population is damaged. It is important to note that within the built environment each building will have significant variations in building geometry, construction quality, maintenance and orientation to wind, with each permutation possessing its own vulnerability. Thus a vulnerability curve due to its empirical nature, describes the average vulnerability of a population of buildings of a similar type, not individual buildings.

The approach starts with the RP curve of the hazard at the local level (at the resolution of the house scale; 25 metres), see Fig. 5.8. For each return-period hazard level, the potential for residential building damage is linked to the hazard through the vulnerability relationship for the residential structure type under consideration (Smith et al 2020). Damage is calculated through the interaction of wind speed, building exposure and vulnerability.

The method to assess cost of damage calculates the 50 to 2000-year return period of loss levels for each building. Losses are then regressed to obtain a Probable Maximum Loss (PML) curve for each building. The Probable Maximum Loss (PML) curve relates damage of a population of similar structures to the likelihood (return-period) for the hazard, see Fig. 5.11.

The losses represented by the curve range from frequent minor losses through to those associated with catastrophic events having disastrous effects on each case study region. Annualised loss, which is evaluated by integrating the area under the PML curve, represents the average annual cost to the region due to exposure to the hazard viewed through a very wide window of time (2000 years adopted). The value of the loss calculated is the full cost of repair or replacement (new for old). This is significantly greater than the insured value where payout is often on a like for like basis (Cechet et al 2012).

5.6 Final note

As in other mathematical models it is important to be aware of the model limitations otherwise incorrect conclusions can be drawn. The main limitations of any model based on extreme value distributions are:

- The results are valid in the limit, an idealised mathematical space of infinite observations, so there is a degree of uncertainty when used with finite samples,
- The short length of records may not be representative of either a location or a region. North Queensland, for instance is struck by cyclones regularly. Most observational datasets of wind speed in the region may not include recent events,
- The model itself is based on observational data that may have large but unknown errors. On the other hand most instruments are calibrated for mean wind speed, extreme gusts records are not calibrated (Jacob 2010),
- The records may not be based on independent observations,
- Extreme value studies are very sensitive to the range of the records selected for the modelling work. Generally the latter part of the records are more reliable (Cechet & Sanabria 2011).

Another important consideration in these types of studies is that the datasets used for extreme value modelling must be homogenous. In the case of wind most datasets comprise winds originated from cyclones, synoptic and convective phenomena. These types of winds are considered separated phenomena, not only because they have dissimilar characteristics but also because they are generated by different physical laws. The extreme value analysis should consider each type of wind independently. This topic however is beyond the scope of this chapter, interested readers should refer to Holmes (2002, 2007), Sanabria & Cechet (2007).

In spite of these limitations, the model discussed in this chapter can be useful to assist engineering, planning and emergency authorities with their decision-making regarding hazardous winds.

5.7 References

AS/NZS 1170.2:2011 Structural Design Actions Part 2 - Wind actions.

BoM (Australian Bureau of Meteorology) (2020). Climate Data Online. http://www.bom.gov.au/climate/data/ Accessed on 19/05/2020.

Cechet R.P. Sanabria L.A. (2011). Australian extreme windspeed baseline climate investigation project. Intercomparison of time series and coincident Dines cup anemometer observations. Geoscience Australia Record 2011/23. GeoCat 71858.

Cechet RP, Sanabria LA, Divi CB, Thomas C, Yang T, Arthur WC, Dunford M, Nadimpalli K, Power L, White CJ, Bennett JC, Corney SP, Holz GK, Grose MR, Gaynor SM and Bindoff NL. (2012). Climate Futures for Tasmania: Severe wind hazard and risk. Technical Report. Geoscience Australia Record 2012/43. GeoCat 74052.

Chen K. (2004). Relative Risks Ratings for Local Government Areas. Risk Frontiers quarterly newsletter, Macquarie University. Vol. 3 issue 3, March 2004.

Coles S. (2001). An Introduction to Statistical Modeling of Extreme Values. Springer series in statistics. London.

Gillelland E. and Katz R.W. (2005a). Extremes Toolkit: Weather and Climate Applications of Extreme Value Statistics. National Center for Atmospheric Research (NCAR). Boulder CO, USA

Gillelland E. and Katz R.W. (2005b). Analysing Seasonal to Interannual Extreme Weather and Climate Variability with the Extremes Toolkit. National Center for Atmospheric Research (NCAR). Boulder CO, USA.

Ginger J. Holmes J.D. and Harper B. (2013). Gust Wind Speeds for Design of Structures. The Eighth Asia-Pacific Conference on Wind Engineering, December 1014, 2013, Chennai, India.

Holmes, J. D. (2002). A reanalysis of recorded extreme wind speeds in Region A. Australian Journal of Structural Engineering, 4, 1.

Holmes, J. D. (2007). Wind Loading of Structures. 2th ed. Taylor and Francis.

Holmes, J. D. Moriarty W.W. (1999). Application of the generalized Pareto distribution to extreme value analysis in wind engineering. Journal of Wind Engineering and Industrial Aerodynamics, 83, 110.

Jacob, D. (2010). Challenges in developing a high-quality surface wind speed data-set for Australia. Australian Meteorological and Oceanographic Journal 60. 227-236

Jagger T.H. and Elsner J.B. (2006). Climatology Models for Extreme Hurricane Winds near the United States. Journal of Climate. Vol. 19, 3220-3236.

Jenkinson, A.F. (1955). The frequency distribution of the annual maximum (or minimum) values of meteorological elements. Quarterly Journal of Royal Meteorological Society, Vol. 81, pp. 58171.

Kjeldsen T.R. Lamb R. Blazkova S.D. (2014). Chapt 8: Uncertainty in Flood Frequency Analysis. In: Applied Uncertainty Analysis For Flood Risk Management. Edited by Beven and Hall. Imperial College Press.

Lechner J.A., Leigh S.D. and Simiu E. (1992). Recent Approaches to Extreme Value Estimation with Application to Wind Speeds. Part I: the Pickands Method. Journal of Wind Eng. and Industrial Aerodynamics, 41-44. 509-519.

Natalini BM, Natalini B, Atencio BA, Zaracho JI (2016) Anlisis de velocidades de viento extremas de 11 estaciones en Argentina perspectivas para una actualizacin del mapa de vientos extremos. Proceedings of the XXIV Jornadas Argentinas de Ingeniera Estructural, Buenos Aires, Sept 28-30

Oztekin T. (2005). Comparison of Parameter Estimation Methods for the Three-parameter Generalized Pareto Distribution. Turk J Agric For. 29 (2005) 419-428.

Palutikof J.P., Brabson B.B., Lister D. H. and Adcock S.T. (1999). A Review of Methods to Calculate Extreme Wind Speeds. Meteorol. Appl. 6, 119132.

Prescott, P., and A. T. Walden (1980), Maximum likelihood estimation of the parameters of the generalized extremevalue distribution, Biometrika, 67, 723 724.

Sanabria, L. A., and Cechet, R.P. (2007). A Statistical Model of Severe Winds, Geoscience Australia Record, Geocat Number 65052.

Sanabria L.A. Cechet R.P. (2015). Improving regional wind hazard assessment by statistical selection for grouping wind station observations. 14th Int. Conference on Wind Engineering (ICWE14). Porto Alegre, Brazil. June 21-26.

Seguro J.V. and Lambert T.W. (2000). Modern Estimation of the Parameters of the Weibull Wind Speed Distribution for Wind Energy Analysis. Journal of Wind Eng. and Industrial Aerodynamics 85 (2000) 75-84.

Stephenson A (2004). A Users Guide to the EVD Package (Version 2.1). Department of Statistics. Macquarie University. Australia.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Smith D.J. Edwards M. et al. (2020). Modelling vulnerability of Australian housing to severe wind events: past and present. Australian Journal of Structural Engineering. Apr 2020. DOI: 10.1080/13287982.2020.1744900

Twan J.A. (1992). Estimating probabilities of extreme sea-levels. Appl. Statistics 41. 77-93.

White CJ, Sanabria LA, Grose MR, Corney SP, Bennett JC, Holz GK, McInnes KL, Cechet RP, Gaynor SM and Bindoff NL (2010). Climate Futures for Tasmania: extreme events technical report, Antarctic Climate and Ecosystems Cooperative Research Centre, Hobart, Tasmania.