Fire Research Note No 1062

INVESTIGATION OF A TECHNIQUE FOR ESTIMATING THE PROBABILITY DISTRIBUTION OF FIRE LOSS BASED ON EXTREME VALUE THEORY

by

F E Rogers

December 1976

FIRE RESEARCH STATION

Fire Research Station
BOREHAMWOOD
Hertfordshire WD6 2BL

Tel: 01 953 6177

Fire Research Note 1062
December 1976

INVESTIGATION OF A TECHNIQUE FOR ESTIMATING THE PROBABILITY DISTRIBUTION OF FIRE LOSS BASED ON EXTREME VALUE THEORY

by

F E Rogers

SUMMARY

The feasibility of applying a model based on the theory of extreme values to the problems of estimating the parameters of the distribution of overall fire loss, and of estimating the annual total loss in fires is investigated.

Results and comprehensive discussion of the results are given for the Textile industry for the years 1966-72. The treatment of the subject is of necessity highly mathematical and somewhat complex in nature but further publications giving details of practical applications are in course of preparation.

Department of the Environment and Fire Offices' Committee
Joint Fire Research Organization

CONTENTS PAGE NO.

Fire Research Note 1062
December 1976

INVESTIGATION OF A TECHNIQUE FOR ESTIMATING THE PROBABILITY DISTRIBUTION OF FIRE LOSS BASED ON EXTREME VALUE THEORY

by

F E Rogers

1. INTRODUCTION

For economic studies, an estimate of total loss in all fires (large and small) is required for each category of industrial and commercial buildings. This information is not available at present. The British Insurance Association publish for each month the national total of fire losses in all occupancies combined, but no breakdown by occupancy is given and their method of estimation is not known.

Now in general, estimates of fire loss are available to the Fire Research Station only for fires in which the total damage to structure and contents was £10,000[⍭] or more and to answer questions in fire protection economics it is necessary to work with these figures. Although the losses in these large fires are of considerable economic importance, the data are inadequate for standard methods of statistical analysis as they represent only a small percentage of the total number of fires. Because of this difficulty, in a series of papers[1,2,3,4] Ramachandran has developed further the statistical theory of extreme values for studying fire losses. In particular, he[3] proposed a model for estimating the values of the location and scale parameters of the probability distribution of fire loss using only the large losses. The main features of the model are reproduced in Section 2. This report is mainly concerned with investigating the feasibility of applying the model to the problem of estimating the annual distribution of fire loss and the annual total loss within a given group of similar fire risks.

⍭ This limit was raised to £15,000 in June 1974, £20,000 in January 1975 and to £25,000 in January 1976.

The Textile industry was chosen for the first application of the model since this industry has one of the highest annual totals of direct loss in large fires of all industrial groups. Additionally, the Textile industry provides more data than any other single industry on large fires in sprinklered buildings. The data for the Textile industry was further classified according to whether the building was single-or multi-storey and whether sprinklers were present or not. This classification was chosen in order that, within each sub-population, the data is reasonably homogeneous, and as a first stage in a cost/benefit analysis of sprinklers. This data is discussed in Section 3.

Various aspects of the proposed model are discussed in Section 4. Subsequently, two methods of using the data with this model are described and discussed in detail. A suitable procedure has been identified which depends upon data availability and the purpose for which the estimates are required. The recommended procedure is stated in a concise form in Section 5.2.

Results are presented and discussed when the procedure is applied to data from each of the above sub-populations of the Textile industry for the years 1966-72. All estimates of total loss in this report refer to direct loss, no account being taken of consequential losses due to loss of production, employment, etc. Further reports will give the results from similar analyses of large fire data in other industries and trades.

Areas where further work is required or would lead to an improvement in the estimation procedure are mentioned briefly. In particular, a third method of estimation was considered but abandoned requiring further development. This is briefly described in Appendix 2.

2. RESUMÉ OF THE THEORY

Since log fire loss has an exponential type distribution, fire loss x may be assumed lognormally distributed, as shown by some actuaries[5,6]. Thus, $z = \log_{10} x$ follows the normal distribution.

The problem is to estimate the mean μ and standard deviation σ of the distribution of z.

Now if
$$z \sim N(\mu, \sigma^2) \tag{2.1}$$
then
$$t = \frac{z-\mu}{\sigma} \sim N(0,1) \tag{2.2}$$

If the observed large fire losses in a year are arranged in decreasing order of magnitude, so that $x_{(m)n}$ is the m^{th} largest observed loss in a sample of size n, then the corresponding values of z and t are

$$z_{(m)n} = \log_{10} x_{(m)n} \tag{2.3}$$
and
$$t_{(m)n} = \frac{z_{(m)n} - \mu}{\sigma} \tag{2.4}$$

Now for sufficiently large n and small m, then over a N year record it can be shown that, for large values of N, $t_{(m)n}$ has the probability density function (p.d.f.)

$$\psi_{(m)}(y_{(m)}) = \frac{m^m}{(m-1)!} e^{-my_{(m)} - me^{-y_{(m)}}} \qquad -\infty < y_{(m)} < +\infty \tag{2.5}$$

where
$$y_{(m)} = A_{(m)n}(t_{(m)n} - B_{(m)n}) \tag{2.6}$$

$A_{(m)n}$, $B_{(m)n}$ are called the extreme value parameters of $t_{(m)n}$ and are given by solving the equations:

$$G(B_{(m)n}) = 1 - \frac{m}{n} \tag{2.7}$$
and
$$A_{(m)n} = \frac{n}{m} g(B_{(m)n}) \tag{2.8}$$

where $g(\omega)$, $G(\omega)$ are respectively the p.d.f. and cumulative distribution function (c.d.f.) of the standard normal (0,1) distribution.

Let the mean and variance of $y_{(m)}$ given in (2.6) be $\bar{y}_{(m)}$ and $\sigma^2_{(m)}$ respectively.

The model which Ramachandran[3] developed to obtain estimates of the parameters μ and σ is given by introducing an error term into equation (2.4) ie $$\bar{Z}_{(m)n} = \mu + \sigma \bar{t}_{(m)n} + \varepsilon_{(m)} \qquad m = 1, \ldots r \tag{2.9}$$

or in matrix terms $$\underline{Z} = C\underline{\theta} + \underline{\varepsilon} \tag{2.10}$$

where $$\underline{Z}' = \left[\bar{Z}_{(1)n} \quad \bar{Z}_{(2)n} \; \ldots \ldots \; \bar{Z}_{(r)n} \right] \tag{2.11}$$

in which $\bar{Z}_{(i)n}$ is the mean over the N samples of the i^{th} ranked observations

and $$C' = \begin{bmatrix} 1 & 1 & \ldots\ldots\ldots & 1 \\ \bar{t}_{(1)n} & \bar{t}_{(2)n} & \ldots\ldots\ldots & \bar{t}_{(r)n} \end{bmatrix} \tag{2.12}$$

where, from (2.6) $$\bar{t}_{(i)n} = B_{(i)n} + \frac{\bar{y}_{(i)}}{A_{(i)n}} \tag{2.13}$$

$$\underline{\theta}' = [\mu \quad \sigma] \quad , \quad \underline{\varepsilon}' = [\varepsilon_{(1)} \quad \varepsilon_{(2)} \; \ldots\ldots \; \varepsilon_{(r)}]$$

A weighted least squares technique is required since the errors do not have constant expectation and are correlated. The weight matrix is considerably simplified by the result obtained by Ramachandran[2,4] that with n and N large and m and l small

$$Cov(y_{(m)}, y_{(l)}) = Var(y_{(m)}) \qquad m > l \tag{2.14}$$

Hence, in matrix terms the estimates required are provided by minimizing the quadratic form

$$(\underline{Z} - C\underline{\theta})' V^{-1} (\underline{Z} - C\underline{\theta}) \tag{2.15}$$

giving $$\underline{\theta} = (C'V^{-1}C)^{-1}(C'V^{-1}\underline{Z}) \tag{2.16}$$

where $$V = \begin{bmatrix} \dfrac{\sigma^2_{(1)}}{A^2_{(1)n}} & \dfrac{\sigma^2_{(2)}}{A_{(1)n}A_{(2)n}} & \ldots\ldots\ldots & \dfrac{\sigma^2_{(r)}}{A_{(1)n}A_{(r)n}} \\ \dfrac{\sigma^2_{(2)}}{A_{(2)n}A_{(1)n}} & \dfrac{\sigma^2_{(2)}}{A^2_{(2)n}} & \ldots\ldots\ldots & \dfrac{\sigma^2_{(r)}}{A_{(2)n}A_{(r)n}} \\ \vdots & \vdots & & \vdots \\ \dfrac{\sigma^2_{(r)}}{A_{(r)n}A_{(1)n}} & \dfrac{\sigma^2_{(r)}}{A_{(r)n}A_{(2)n}} & \ldots\ldots\ldots & \dfrac{\sigma^2_{(r)}}{A^2_{(r)n}} \end{bmatrix} \tag{2.17}$$

Thus, it is possible to estimate μ and σ based on the combined observations of the r largest extremes from each of N large samples, and the only data required are

(1) The vector of observations, $\underline{\underline{Z}}$

(2) The sample size, n

(3) The values $\bar{y}_{(i)}$, $\sigma^2_{(i)}$ which have been tabulated by Ramachandran[2]

(4) The values of the extreme value parameters $A_{(i)n}$, $B_{(i)n}$ which are constants for a given sample size, and are easily calculated from tables of the standard normal distribution.

A full derivation of the above results is given in Ramachandran[3].

3. THE DATA

3.1 The observed large losses

The Fire Research Station (FRS) receives reports of all fires attended by the local authority fire brigades. Also, the Fire Protection Association (FPA) supply FRS with particulars of fires in which the total direct loss is in excess of £10,000≠. These two sources of information are collated at FRS where there is a data-bank of large fire records dating back to 1965.

For a given population, ie in this application the population consists of all fires in the Textile industry, the minimum information required of each large fire for the purposes of estimating the parameters of the distribution of fire loss by the technique described in this report is:

(i) the total direct loss resulting from the fire

(ii) the number of floors in the building, a basement counting as an extra floor

(iii) the effect of sprinklers if these were present. If a sprinkler system was present but did not operate, then these observations were omitted from both the sprinklered and the non-sprinklered sub-populations in order to avoid bias in the data - sprinklers tend to be installed in those buildings which have a greater value at risk and in which a higher frequency of fire is expected.

≠ See footnote on p 1.

If the computer files showed the number of floors, or effect of sprinklers as being unknown, it was possible, for some large fires, to obtain sufficient information from the microfilm record of the original report to be able to classify the loss into one of the sub-populations, eg description of damage may indicate building is multi-storeyed. Even so, there were some large losses which could not be classified into any of the sub-populations. To obtain an overall estimate of the total loss in the Textile industry, the unclassified large losses were tabulated separately, and then the total unclassified large fire loss was simply added to the total losses obtained for the other sub-populations.

Reliable information on large losses, together with information on sprinklers, and other particulars of buildings involved in large fires is available only for 1966 and later years. After extensive checking, tables of raw data for the Textile industry have been compiled for each of the sub-populations for the period 1966–72, and these are given in Tables 1 to 5, Appendix 1. Each year's data is tabulated in descending order of magnitude and the largest loss in each year has been assigned rank 1. The rank, within years, of each observation is given in column 1 of the above tables.

3.2 Transforming the data

For any estimates based on observations over a number of years, it is necessary to adjust the observations to allow for inflation. Whenever necessary, the losses have been deflated to 1966 price levels using factors based on the Retail Price Index – see Table 6, Appendix 1.

Additionally, as a consequence of previous work by Ramachandran[1,4] it is assumed that the parent probability distributions of the sub-populations are lognormal. This assumption has also been made by many actuaries working in the field of fire losses, eg Benckert[5], Ferrara[6]. For various reasons, which will be given later, it is better to work with the normal distribution on a log scale. The loss must be positive, and has no upper limit; therefore the transformation

$$Z = \log_{10}(x) \tag{3.2.1}$$

is appropriate, where x is the observed loss, adjusted for inflation.

3.3 Estimation of the parent sample size

An important part is played by the parent sample size in the estimation procedure. In the context of this application, the basic parent sample size is the number of fires each year in each sub-population. These values were estimated thus:

(i) From the UK Fire and Loss Statistics[10] obtain the total number of fires each year in the Textile industry.

(ii) From computer records obtain the observed relative frequency of fires in each of the four sub-populations, based on a four-year period, say.

(iii) According to a survey conducted some years ago by FRS, about one-third of fires in sprinklered buildings are neither attended by fire brigades, nor reported to the Joint Fire Research Organisation[9]. Therefore, the observed relative frequency for each of the sprinklered sub-populations was increased by 50 per cent.

(iv) The normal distribution has a monotonic increasing failure rate function and yet, from Ramachandran[1], log fire loss has conceptually a U-shaped failure rate function. Thus the assumption that log fire loss is normally distributed is valid only for the range of values which corresponds to fires having developed beyond the 'infant mortality' stage. From the UK Fire and Loss Statistics[10] it can be seen that approximately 50 per cent of fires in the Textile industry are small in that they are confined to exterior components, common service spaces and appliances. These small fires were excluded from the total number of fires to allow for infant mortality and the observed relative frequencies reduced accordingly.

The relative frequencies were then applied to the total number of fires each year in the Textile industry to produce annual estimates of the number of fires in each of the four sub-populations. The results are given in Table 7, Appendix 1.

4. THEORETICAL ASPECTS OF THE ESTIMATION PROCEDURE

4.1 Examination of the variance-covariance matrix V

Using the asymptotic values of $\bar{y}_{(m)}$ and $\sigma^2_{(m)}$ as given by Ramachandran[2], estimates of the location and scale parameters μ and σ were obtained from data in the period 1966–71 (later extended to 1966–72) for each of the four sub-populations based successively on r = 3, 4, ... R extremes, abbreviated to $r = 3(1)R$, where R depends on the amount of data available in each sub-population. In addition, the associated statistics residual variance, var (μ), var (σ), and cov (μ, σ) were also calculated.

Now the form of the V matrix given in (2.17) depends upon the result (2.14) which is an asymptotic result being true only for large values of n and N and small values of m and l . Also it is known that the distribution of the extreme values from a normal parent distribution converge slowly with increasing n to the asymptotic form of the extreme value distribution. However, for the sprinklered fires in single-storey buildings sub-population, from Table 7, Appendix 1, the average sample size is approximately 80. Furthermore, the maximum number of years (samples) N which will be used in the estimation procedure (in this application) is 7, and yet the asymptotic results require N large. Is it justified in this situation to use the asymptotic values of $\bar{y}_{(m)}$ and $\sigma^2_{(m)}$ and the asymptotic result (2.14)? There are no published results available at present to answer this question.

Two alternative forms of the variance-covariance matrix were therefore considered in detail in an attempt to make the model more realistic in its application to the data for the Textile industry. However, these alternatives were both finally rejected in favour of the matrix obtained by using the asymptotic values of $\bar{y}_{(m)}$ and $\sigma^2_{(m)}$.

Firstly, to allow for small values of N, small sample estimates of $\bar{y}_{(m)}$ and $\sigma^2_{(m)}$ were calculated from the $j/(N + 1)$ quantile points of the p.d.f. of $\bar{y}_{(m)}$, for $m = 1, \ldots r$, $j = 1, \ldots N$. The estimates of $\bar{y}_{(m)}$, and $\sigma^2_{(m)}$, were then substituted for the asymptotic values of these statistics wherever they appeared in the matrices C and V. The residual variance obtained when these alternative estimates of $\bar{y}_{(m)}$ and $\sigma^2_{(m)}$ were used, were always greater than when they took their asymptotic values with the same set of data. Also, since the quantile points of $\psi_{(m)}(y_{(m)})$ have not been tabulated (except for $m = 1$), this modification leads to a much less efficient computational procedure.

A second alternative, to allow for small values of N and n is to again use the small sample estimates of $\bar{y}_{(m)}$, and $\sigma^2_{(m)}$, but furthermore to calculate a small sample estimate of cov($y_{(m)}, y_{(l)}$). A simple estimate is given by associating the j^{th} ranked observation within the m^{th} ranked extreme of each sample with the $j/(N + 1)$ quantile point of $\psi_{(m)}(y_{(m)})$. This method gave very poor results since the V matrix so obtained was almost always ill-conditioned.

The modified estimates of $\bar{y}_{(m)}$, $\sigma^2_{(m)}$, and cov($y_{(m)}, y_{(l)}$) mentioned here are only two of a large number of alternatives. There was no evidence offered by the two modifications considered here in detail to suggest that any improvements to the estimation are possible by considering further modifications. Hence, it is concluded that the best estimates of μ and σ are obtained from equation (2.16) when the C and V matrices (equations (2.12) and (2.17) respectively) are calculated using the asymptotic values of $\bar{y}_{(m)}$, and $\sigma^2_{(m)}$. This statement is in agreement with Ramachandran's[3] heuristic result and is made subject to the proviso that $n \gg 50$ in order that the application of extreme value theory is in fact appropriate.

4.2 Further comment on the asymptotic form of the V matrix

Before finalising the technique to be used to estimate the annual total fire loss in a given sub-population, a few further comments are made on the use of the asymptotic values of $\bar{y}_{(m)}$ and $\sigma^2_{(m)}$ in the V matrix:

(i) The form of the V matrix using the asymptotic values of the mean and variance of $y_{(m)}$ is to be used in all subsequent applications of the estimation procedure. How robust then is V^{-1}, and the results obtained using this form of V ? (V^{-1} is in general a tridiagonal matrix with the non-zero elements generally in the range $|v_{ij}| < 100$. The $(i, j)^{th}$ element will be +ve for $i+j$ even, and -ve for $i+j$ odd).

Using the data for the period 1966-71 for the non-sprinklered, multi-storeyed sub-population, the inflation factor was changed from 131.7 to 126.0. This will have the effect of perturbing each element of the Z vector by the amount k where

$$k = -\frac{1}{6}\left\{\log_{10} 131{\cdot}7 - \log_{10} 126{\cdot}0\right\} = -0{\cdot}0032 \quad (4.2.1)$$

This perturbation is reflected exactly in the results, with the estimated value of μ being reduced by the amount k, and σ remaining unchanged, even when 20 extremes are used so that V is a 20 x 20 matrix. Thus, this form of V will be regarded as being robust when up to 20 extremes are being used. It is not known how many more extremes may be included.

(ii) The C and V matrices as defined in Section 2 are fixed for a given parent distribution and sample size. How will an error in the estimated parent sample size affect the results of the estimation procedure? A brief investigation of the effect of changing the sample size was carried out. For very large samples, an error in the estimated sample size has little effect on the estimated values of μ and σ. However, for small values of n errors in the estimated sample size are reflected in the estimates of μ and σ, but also since n is small it is likely that there will only be a small number of extreme observations. In this situation, unless the estimated sample size is grossly in error, the effect of any errors in estimating the sample size will be small in comparison with the effect of having only a few observations to work with. Ramachandran[3] has shown that a variation in sample size does not seriously affect the value of $A_{(m)n}$, but $B_{(m)n}$ in (2.7) is affected approximately according to the following formula

$$B_{(m)n_j} = B_{(m)n} + \frac{1}{A_{(m)n}} \log_e\left(\frac{n_j}{n}\right)$$

where n may be regarded as the true sample size and n_j the estimated value. $A_{(m)n}$, $B_{(m)n}$ generally take values in the range (1,4). Thus only large values of the ratio (n_j/n) would have a serious effect on the value of $B_{(m)n}$. The parameters μ and σ are functions of $A_{(m)n}$ and $B_{(m)n}$.

(iii) Estimates of μ and σ can be obtained using the observations from one extreme only. Thus if R extremes are known, then R estimates of μ and σ can be obtained. The method of moments, leads to the following estimates

$$\sigma = \frac{A_{(m)n}}{a_m} \qquad (4.2.2)$$

$$\mu = b_m - \sigma B_{(m)n} \qquad (4.2.3)$$

where $B_{(m)n}$, $A_{(m)n}$ are given by equations (2.7) and (2.8) respectively

and
$$a_m = \frac{\sigma_m}{\sigma_{mz}} \qquad (4.2.4)$$

where
$$\sigma_{mz}^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left(Z_{(m)n.i} - \bar{Z}_{(m)n} \right)^2 \qquad (4.2.5)$$

and
$$b_m = \bar{Z}_{(m)n} - \frac{\bar{y}_{(m)}}{a_m} \qquad (4.2.6)$$

where $\bar{y}_{(m)}$, $\sigma_{(m)}^2$ are the mean and variance of $y_{(m)}$.

While proposing the method mentioned above, Ramachandran[3] pointed out that it is difficult to draw reliable conclusions from estimates based on just one extreme. These estimates were calculated for each sub-population for the period 1966–71, for each available extreme. As an example the results obtained for the non-sprinklered single-storey sub-population are given in Table 8, Appendix 1. These confirm that estimates obtained from observations on one extreme only vary widely. For practical purposes estimates of the parameters based on data taken from several extremes are required.

(iv) The distribution of the error is not known. If, as a first approximation, the errors are assumed to be multi-variate normally distributed, then it is possible to obtain a $(1-\alpha)\%$ joint confidence region for $\hat{\mu}$, $\hat{\delta}$ from

$$(\hat{\underline{\theta}} - \underline{\theta})' C' V^{-1} C (\hat{\underline{\theta}} - \underline{\theta}) = \frac{2}{R-2} (\underline{Z}' V^{-1} \underline{Z} - \hat{\underline{\theta}}' C' V^{-1} \underline{Z}) F_{2,R-2,1-\alpha} \qquad (4.2.7)$$

where $\hat{\underline{\theta}}$ is the estimated parameter vector $= \left[\hat{\mu} \quad \hat{\delta} \right]$

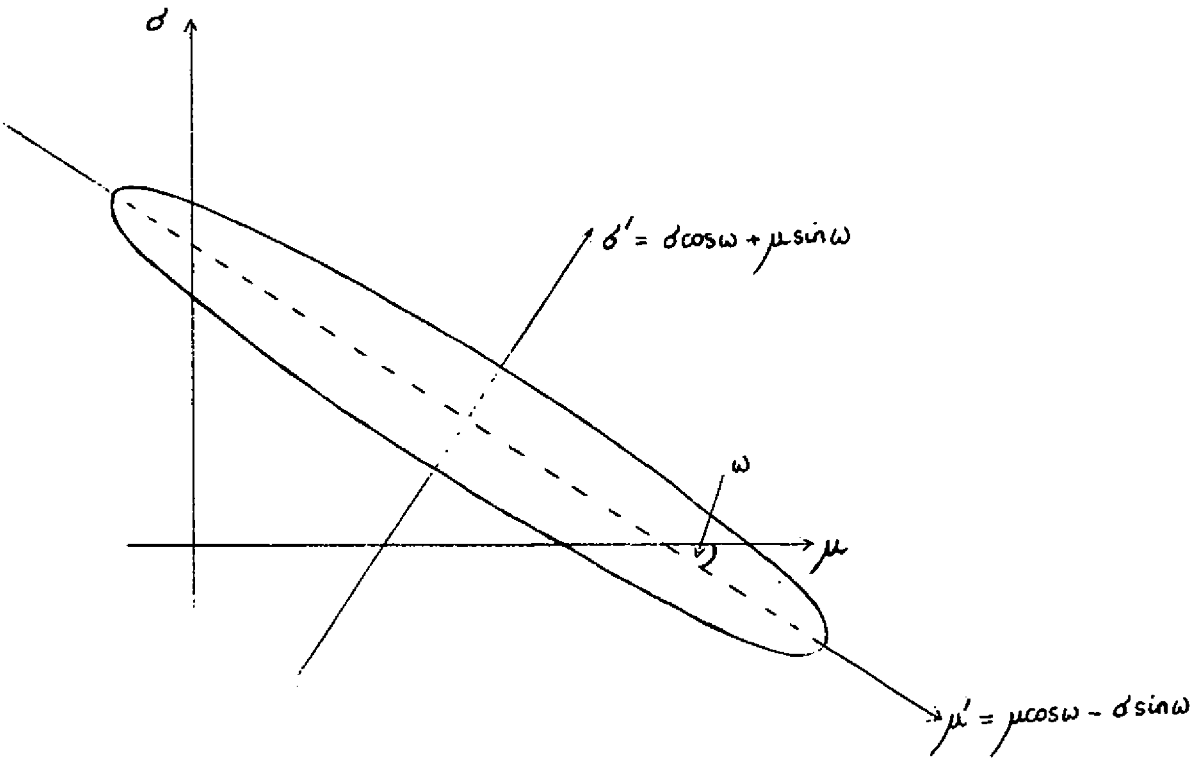
(4.2.7) is a quadratic expression in (μ, δ).

The main objective in calculating the confidence region was to assess the estimates provided by the method of moments applied to data from individual extremes compared with the composite estimates obtained by considering all extremes taken together. The estimates based on individual extremes have subsequently been rejected - see 4.2 (iii). Additionally, the confidence region provided is only an initial estimate based on an approximate assumption that the errors are multi-variate normally distributed.

However, a 95 per cent confidence region was obtained for several of the estimations and the following general comments can be made for each of the sub-populations:

(a) Each confidence region is a real ellipse with the ratio of the lengths of the major axis to the minor axis being at least 7 : 1.

(b) The major axis forms an angle ω with the μ axis, and ω is generally in the range -25° to -35°, eg



(c) $\hat{\underline{\theta}}$ is the centre of the ellipse

(d) There is dependence between $\hat{\mu}$ and $\hat{\sigma}$, and in fact they have nearly perfect negative correlation

(e) $\hat{\sigma}'$ is well determined (short minor axis)

$\hat{\mu}'$ is not well determined (long major axis)

It is difficult to interpret the confidence region beyond the general conclusion stated above.

5. USE OF THE ESTIMATION PROCEDURE TO OBTAIN ANNUAL ESTIMATES OF TOTAL LOSS

5.1 The techniques

Let the location and scale parameters of the transformed distribution of fire loss associated with year i be denoted by μ_i and σ_i respectively. The data has been used in two ways to provide estimates of these parameters. These are:

METHOD 1. Obtain the composite estimates, denoted by μ_c , σ_c , for a N year period by including data from each year in the period and calculating the Z vector as specified in equation (2.11). Now an assumption basic to the extreme value distributions is that each sample is drawn independently from the same population, ie the annual distributions of fire loss are assumed to be independently and identically distributed.

Hence
$$\mu_i = \mu_c + k_i$$
$$\sigma_i = \sigma_c \qquad i = 1, \dots N \qquad (5.1.1)$$

where $k_i = \log_{10}$ (inflation factor for year i)

METHOD 2. Follow Lloyds[7] suggestion and let $\bar{t}_{\omega n}$ in equation (2.13) be replaced by $t_{\omega n}$, the single observation forming the i^{th} largest loss in a sample of size n , transformed by equation (2.3) and normalized by equation (2.4).

This is equivalent to Method 1 with N = 1. This method leads directly to estimates of μ_i and σ_i and no adjustment for inflation is necessary.

Certain situations with regard to data availability definitely favour the application of one or other of these two methods. In other cases neither method is distinctly better than the other, and the final choice of estimates requires a certain amount of judgement. The choice of estimates is also influenced by their intended use. A full discussion is given in Section 5.2. A third method was also considered but was abandoned requiring much further development. This third method is briefly described with suggestions for further work in Appendix 2.

There are two ways in which the annual estimates μ_i , and σ_i obtained by either of the above methods, can be used to give an annual estimate of the total loss.

(i) Direct method

Obtain $\bar{x}_i$, the arithmetic mean for the full range of fire loss, in terms of the original units, using

$$\bar{x}_i = \exp\left\{\frac{\mu_i}{T} + \frac{\sigma_i^2}{2T^2}\right\} \qquad (5.1.2)$$

where T = $\log_{10}e$

(The factor T is required since the observations were transformed using logs to base 10, rather than natural logs)

Then Total loss in year i = $\bar{x}_i \, . \, n_i$ (5.1.3)

where n_i = estimated number of fires in year i .

(ii) Indirect method

By inspection of the data, estimate the lower bound of the known large losses for each year (the lower bound is not necessarily £10,000). Denote this lower bound by b_o . Now estimate the arithmetic mean of the distribution of fire loss, in terms of the original units, over the restricted range (a_o, b_o) – a_o has been assigned the value £25 throughout this report such that the range (0, 0.025) is excluded when calculating the restricted means, $\bar{x}_{Ri}$, to allow for 'infant mortality' fires. Then,

Total loss for year i = $\bar{x}_{Ri}$. (n_i – No. of known large fires)

+ sum of known large losses (5.1.4)

It can be shown that if $Z = \log_{10}x \sim N(\mu, \sigma^2)$ then

$$\bar{x}_{Ri} = \bar{x}_{(a_o, b_o)} = \frac{1}{T}.\exp\left\{\frac{\mu}{T} + \frac{\sigma^2}{2T^2}\right\}.\left\{G(b_1) - G(a_1)\right\} \qquad (5.1.5)$$

in which $$G(\omega) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\omega} \exp\left(-\frac{t^2}{2}\right) dt$$

and $$a_1 = \frac{\log_{10}a_o - \mu}{\sigma} - \frac{\sigma}{T}$$

$$b_1 = \frac{\log_{10}b_o - \mu}{\sigma} - \frac{\sigma}{T}$$

5.2 Discussion of the techniques

Both Methods 1 and 2 were applied to a large number of permutations of the data given in Tables 1-4, Appendix 1. For either method, if R is the maximum number of observations available in a sub-population for a particular year or period of years, then estimates were obtained based on the ..... R - 3, R - 2, R - 1, R largest observations taken together to see how the estimates changed as more data were incorporated into the estimation. Ideally the estimates of μ and σ should converge to their true values as additional observations are used. However, no test of convergence is available at this stage, but in general it is self-evident whether the estimates are reliable since as further data is included in the estimation the estimates either remain fairly constant or vary considerably. The value of R required to give stable estimates of μ and σ is not known, but Method 2 certainly requires R to be larger than if Method 1 is used. As a guideline, based purely on examination of the results obtained for the Textile industry, R should form approximately 5 per cent of the sample when Method 1 is used but 8 per cent when Method 2 is used. It is not known if these quantities are true in general.

For many of the estimates of μ and σ annual estimates of total loss were also calculated by both the direct and the indirect methods given in Section 5.1. It was soon established that the indirect method was to be preferred at all times. However, the choice of the best estimates of μ and σ for a particular year is less clear. Whether Method 1 or Method 2 should be used depends on several factors, including the purpose for which the estimates are required, and data availability. Some of the main advantages and disadvantages of each method are given below, and although it has not been possible to give a firm procedure for establishing the best estimates of μ and σ, good guidelines have been identified.

In Method 1, the Z vector (equation (2.11)) is obtained by averaging observations over a N year period. Hence there is an implied assumption that the annual distribution of fire loss remains unchanged except for the effect of inflation, over the N year period. This assumption is used in order to obtain annual estimates of μ and σ. Changes in the data may be due either to a change in the parent population or simply random variations. Unfortunately, it is not possible to test for a change in the parent distribution, but over a seven year period as considered in this application, only mild deviations would be expected. However, if the annual distribution of fire loss really is changing then a basic assumption has been violated and the annual estimates of μ and σ obtained by Method 1 will be poor. This can be seen by comparing the observations for a given year with the distribution of fire loss estimated by Method 1 for that year. Even so, Method 1 will still lead to an estimated distribution which is typical for the sub-population under consideration and so may be used in any problem requiring a long-term estimate of the distribution of fire loss, eg cost/benefit studies.

Although Method 1 is quick to apply, practically it has a few short-comings. In particular, Method 1 will certainly not make use of all the available data, eg in the sprinklered multi-storey sub-population, 1968 and 1971 each have 16 known large losses, but the Method 1 estimate over years 1966–72 would only use extremes 1 to 10 and ignore the remainder. Also, the sample size used in any Method 1 estimate is an average of the numbers in the individual samples which contribute to the Z-vector. Thus, Method 1 does, to some extent, ignore certain characteristics of the individual years which are covered by Method 2.

Method 2 is also quick to apply and has certain advantages over Method 1 if annual estimates are required. Fuller use of the data is made by Method 2, allowing all observations in a year to be included in the estimation. In addition, Method 2 uses the sample size specific to the year under consideration. Also no adjustment for inflation is necessary for this method. Thus Method 2 leads directly to annual estimates at current prices using observations from individual years. Therefore this method is not affected by a parent population which may be changing with time, and in fact may show up any such changes.

The estimated annual distribution of fire loss by Method 2 was in general in closer agreement with the data than the distribution given by Method 1. Method 2 is also much simpler to interpret and frequently variations in the estimates, as more data is included in the estimation, may be traced back to inconsistencies in the data. Note, however, that the Z -vector for M ethod 2 is made up of individual observations, and so each element of $\underline{Z}$ for Method 2 will have a larger sampling variation than the corresponding $\underline{Z}$ for Method 1. This leads to Method 2 being more sensitive to changes in the data. This may have such an effect that even if a large amount of data is available, the occurrence of one or two very large losses may lead to very poor estimates of μ and σ by Method 2. An attempt was made to overcome the problem of heavy-tailed data by obtaining an estimate based on observations ranked k, k + 1, k + 2 .... with k > 1. Further work is required before the estimates given by this modification may be fully assessed; therefore, its application cannot be recommended at this stage. This same modification when used with Method 1 led to estimates which varied little from those obtained when all ranks were included. This shows Method 1 to be much less sensitive to heavy-tailed data - as is to be expected.

The above discussion indicates that the choice of Method 1 or Method 2 estimates is straightforward. If annual estimates are required, use Method 2 estimates when these are stable, otherwise use Method 1 estimates adjusted to the correct price level. Method 1 estimates must be used for problems requiring long-term estimates. However, when there is limited data available, both Method 1 and Method 2 estimates obtained in the manner described in Section 5.1 may be unreliable. In this case the data may be used in a different way by pooling within differing groups of years, eg adjust for inflation then pool and rank data within the three periods 1966–7, 1968–9, 1970–1. Either of Methods 1 or 2 may then be used on the rearranged data. The fewer observations that are available, the longer the period over which the data must be pooled. In some cases this may result in the data from the entire period being pooled into one vector of observations in which case only Method 2 is applicable, but in this case the estimates may be considered as long-term estimates. Obviously there are many ways in which the available data may be pooled but there is no rigorous test available for choosing between the estimates which these yield. In most cases it was possible to make a choice by comparing the probability of observing a loss in excess of any given value, eg £10,000 or £100,000, for each of several estimated distributions of fire loss with the same probability based on the data. Alternatively, the reduced values $y_{(m)}$ corresponding to $Z_{(m)n}$ may be calculated for several estimated values of the parameters μ and δ . These may then be compared since $2me^{-y_{(m)}} \sim \chi^2_{2m}$. However, this procedure is tedious and since the reduced values may only be compared for each rank individually, a decision is not always possible by this method.

The following guidelines should be noted when using the estimation procedure discussed in this report:

(i) Estimates of μ , δ by Method 2 are to be preferred in any problem requiring an annual estimate of the distribution of fire loss. If the data is not heavy-tailed, and the estimates μ , δ given by Method 2 appear to be converging as more information is included in the estimation, then the Method 2 estimates should be used to give the annual distribution of fire loss. Note, however, that no test has been used for heavy-tailed data and at present there is no test of convergence available.

(ii) If in a problem requiring an estimate of the annual distribution of fire loss the Method 2 estimates are not suitable, then the Method 1 estimate adjusted to the correct price level may be used to obtain an annual estimate of total loss. Method 1 estimates of μ and σ should not, however, be considered as specific to any individual year.

(iii) For a general distribution of fire loss for any specific sub-population the Method 1 estimates of μ and σ should be used.

(iv) For any estimate of total loss the indirect method of using the estimated parameters of the distribution of fire loss should always be used, regardless of the method of estimating μ and σ.

(v) In problems with very limited data available the data should be rearranged in a variety of ways and estimates obtained by both Methods 1 and 2. The choice of estimates to be used is then made essentially on an empirical basis. Occasionally this may lead to the Method 2 estimates of μ and σ being used as long-term estimates whenever the data must be pooled over all years.

5.3 Further comment on the estimation procedure

It is important to know where errors may enter the estimation and how these may affect the estimates of μ and σ and hence the expected loss. A detailed examination was carried out of a set of results which are typical for the problem dealt with in this report. The following statements have general validity over the range of values found in estimating the distribution of fire loss.

(i) Let μ_E be the estimated value of the location parameter based on an estimated sample size n_E. If the true values are μ_T, n_T respectively then it can be shown theoretically that

$$\mu_T - \mu_E \simeq -k \log_e\left(\frac{n_T}{n_E}\right) \qquad (5.3.1)$$

in which k depends on the number of extremes used in the estimation, and the scale parameter and intensity function of the distribution being estimated. Formula (5.3.1) is based on a model given in Ramachandran[3] for the parameter $B_{(m)n}$. The computation of k is difficult.

A sequence of estimates of μ and δ was obtained from a single set of data but allowing the sample size to vary. From these results an estimate of k = +0.85 was obtained with a high correlation.

Thus
$$\mu_T - \mu_E \simeq -0.85 \log_e\left(\frac{n_T}{n_E}\right) \tag{5.3.2}$$

For large values of n_T and

$$\left| 1 - \frac{n_T}{n_E} \right| \leq \delta \qquad , \quad \delta \text{ small}$$

then (5.3.2) may be approximated by

$$\mu_T - \mu_E \simeq -0.85\left[1 - \frac{n_T}{n_E}\right] \tag{5.3.3}$$

Equation (5.3.2) shows that

$$|\mu_T - \mu_E| \leqslant 0.2 \quad \text{for} \quad |\delta| \leqslant 0.2$$

or
$$0.8 \leqslant \frac{n_T}{n_E} \leqslant 1.2$$

Thus μ_E is not sensitive to moderate errors in the estimated value for n_T.

(ii) Since the estimated values of μ and δ have nearly perfect (negative) correlation

$$\delta_E = a + b\mu_E \tag{5.3.4}$$

and the parameters a and b may be well determined for individual sub-populations.

(iii) Let $\bar{x}_E$ be the expected loss based on the estimated values μ_E and δ_E,

then
$$\bar{x}_E \doteq \exp\left\{\frac{\mu_E}{T} + \frac{\delta_E^2}{2T^2}\right\}$$

in which $T = \log_{10}e$ is necessary since logs to base 10 were used in estimating μ_E, δ_E.

As a result of (5.3.4)

$$\bar{x}_E = \exp\left\{ f(\mu_E) \right\}$$

where $$f(\mu_E) = \frac{\mu_E}{T} + \frac{(a+b\mu_E)^2}{2T^2}$$

Similarly $$\bar{x}_T = \exp\left\{ f(\mu_T) \right\}$$

so that $$\frac{\bar{x}_T}{\bar{x}_E} = \exp\left\{ f(\mu_T) - f(\mu_E) \right\} = \exp\left\{ h(\varepsilon) \right\}$$

where $$\varepsilon = \mu_T - \mu_E$$

$h(\varepsilon)$ is a quadratic function with one root at zero. The function $h(\varepsilon)$ was determined for μ and σ as estimated for the sprinklered single-storey sub-population of the Textile industry. ε was given a series of values and $\exp\left\{h(\varepsilon)\right\}$ evaluated. Table 5.3 given below shows that no serious error would occur for $|\varepsilon| \leqslant 0.2$.

$\varepsilon = \mu_T - \mu_E$	– 0.2	– 0.1	0.0	0.1	0.2
$\frac{\bar{x}_T}{\bar{x}_E}$	1.07	1.03	1.00	0.98	0.98

Table 5.3 Approximate ratio of true expected loss to estimated expected loss when the true value of μ differs from the estimated value cf μ by the amount ε in the sprinklered single-storey sub-population of the Textile industry.

Now note that

a) The estimates of μ and σ given in Table 13, Appendix 1 are considered to have converged to their true values. No rigorous test for convergence was possible but an estimate of μ would certainly be unacceptable if successive estimates of μ, obtained as each additional observation was included in the data, varied by an amount greater than 0.2.

b) From (i) the estimate of expected loss is not sensitive to moderate errors in the estimated sample size.

(iv) The cumulative distribution functions for combinations of μ_ε, σ_ε related by equation (5.3.4) intersect at $x = 10^{b/a}$. The intersect will generally occur at $x \geqslant$ £10,000 in the specific problem studied in this report. The value of the c.d.f. at $x =$ £10,000 is of particular interest since this value is used to calculate the expected loss by the indirect method described in Section 5.1. Again errors in μ of the order $|\mathcal{E}| \leqslant 0.2$ have little effect on this quantity.

Overall then no serious errors will occur in the estimation and use of μ and σ as described in this report.

5.4 Discussion of results for the Textile industry

The Textile industry as defined in the Standard Industrial Classification[11] includes the production of man-made fibres, the processing of cotton and woollen fibres, and the manufacture of carpets, ropes, knitted goods, and made-up textiles etc.

During the period 1966–72 there were 501 large fires recorded in the Textile industry resulting in a total large fire loss of almost £42M at 1966 prices. There were 5 losses in excess of £1M and a further 10 losses in the range £0.5–£1.0M at current prices. There were 5 large fires in 1969 with a loss in excess of £0.8M, whereas no fire in 1970 incurred a loss in excess of £0.4M.

After adjusting for non-reported fires in the sprinklered sub-populations and excluding infant mortality fires in all sub-populations (see Section 3.3), the estimated total number of fires in the Textiles industry during the period 1966–72 was 4,880, an average of nearly 700 per annum. Thus the probability of observing a large fire, given a fire which has developed beyond the infant mortality stage is, for the Textile industry as a whole, approximately 0.10 based on the raw data.

(i) Sprinklered single-storey buildings

For this sub-population there were only 37 large losses recorded during the period 1966–72. As mentioned earlier, estimates were obtained by using the data in three ways:

(a) Pooling the data into two groups corresponding to the periods 1966-1968, and 1969–1972, then using Method 2 on each group separately - results labelled Method 2A, Table 9, Appendix 1.

(b) Pooling all data into one group for the period 1966–72 and then using Method 2 - results labelled Method 2B, Table 9, Appendix 1.

(c) Use Method 1 on the data pooled into two groups as for (a) - results labelled Method 1, Table 9, Appendix 1.

The results obtained by these three methods were close, and there is no real difference between them. However, neither Method 1 nor Method 2B results can reflect any changes over the years in the parent distribution. Hence, the results from Method 2A are taken as being the best available for this sub-population. Estimates of annual total loss for this sub-population are given in Table 14, Appendix 1. A summary is given in Table 15, Appendix 1 from which it can be seen that the total loss in sprinklered fires in single-storey buildings in the Textile industry during the period 1966–72 was approximately £2.11M at an average of £0.30M per annum or £3,657 per fire at 1966 price levels.

The estimates of μ and σ by Methods 1 and 2B lead to good agreement between the data for the full period 1966-72 and the fitted distribution. Thus either of these sets of estimates may be used in any problem requiring a 'long-term' estimate of the distribution of fire loss for this sub-population. Therefore, the probability of observing a large loss, given a fire which has developed beyond the infant mortality stage, is approximately 0.06 based on the Method 1 estimates - which is in agreement with the value obtained from the raw data.

(ii) Sprinklered multi-storey buildings

For this sub-population estimates were obtained by Methods 1 and 2, as described earlier. In this sub-population there is a wide variation in the data, both between years and within years. This led to varying and unstable estimates of μ and σ by Method 2. Hence, the Method 1 estimates are taken to be the best available for this sub-population.

The estimates of annual total fire loss in the sprinklered multi-storey sub-population of the Textile industry are given in Table 14, Appendix 1, and summarized for the period 1966-72 in Table 15, Appendix 1. This shows the total loss in the period reported to be approximately £8.73M at an average of £1.247M per annum or £4,037 per fire at 1966 prices.

The sprinklered multi-storey sub-population is by far the largest under consideration in this report having more than three times the number of fires recorded in the sprinklered single-storey sub-population and more than double the number recorded in each of the two non-sprinklered sub-populations. However, these ratios are not maintained in terms of the number of large fires, and in fact, this sub-population has a smaller probability of observing a large fire than any of the other sub-populations, being approximately 0.04.

It is not known why the sprinklered multi-storey sub-population has a smaller probability of observing a large fire than the sprinklered single-storey sub-population. Possible reasons are that the various activities carried on in the multi-storey buildings may on the whole present a smaller fire hazard, or that the multi-storey buildings may have more compartments etc.

(iii) Non-sprinklered single-storey buildings

Again, estimates were obtained by each of Methods 1 and 2. The Method 2 estimates were stable and have been used to provide estimates of annual total loss in the non-sprinklered single-storey sub-population of the Textile industry for each of the years 1966-72. These results are given in Table 14, Appendix 1 and are summarized in Table 15, Appendix 1, from which the total loss during the period 1966-72 was estimated at £7.60M at an average of £1.09M per annum or £7,559 per fire at 1966 prices.

The Method 1 estimates were also stable and therefore may be used in any problem requiring a 'long-term' estimate of the distribution of fire loss in this sub-population. The probability of observing a large loss in this sub-population of the Textile industry, given a fire which has developed beyond the infant mortality stage, is approximately 0.10 based on the Method 1 estimates. This is in exact agreement with the probability estimated from the raw data. Probabilities for other levels of loss may be obtained from Fig. 2.

(iv) Non-sprinklered multi-storey buildings

Estimates of the parameters of the distribution of fire loss in this sub-population were obtained by Methods 1 and 2. Again, the Method 1 composite estimates and Method 2 estimates for each year's data, taken separately, were found to be stable. However, the probability of observing a large loss given a fire which has developed beyond the infant mortality stage, based on the estimated values of the parameters of the distribution of fire loss, is, in general, in poor agreement with the same probability based on the raw data.

A histogram was constructed for the entire data given in Table 4, Appendix 1, from which it appears that the assumption that the distribution of log loss is normal, is not valid for this sub-population. There is no simple test of this assumption. However, the author found that in subsequent applications of this method of estimation to other industries and distributive trades[12], this assumption was reasonable and did not appear to be violated. Also, according to actuaries in the field of fire insurance[5,6], the lognormal distribution shows a good fit especially for large values of the variable.

The probable explanation of the failure of this assumption with this sub-population is that the sub-population is still heterogeneous - the processes carried out in the Textile industry are many and varied, such that the sub-population of non-sprinklered multi-storey buildings in the Textile industry does not form a group of occupancies with similar fire risks. It is not known how this sub-population should be further sub-divided to obtain two or three more homogeneous sub-populations, or if the information is available so that this may be done. Hence, the estimates obtained for the non-sprinklered multi-storey sub-population should be considered to be initial estimates which may be improved at a later date by further sub-division of this sub-population.

The Method 2 estimates have been used to obtain annual estimates of total loss and these are given in Table 14, Appendix 1. The summary of these results shows the total loss in this sub-population during the period 1966-72 to be approximately £29.11M at an average of £4.16M per annum or £26,655 per fire at 1966 prices. Until improved estimates become available the Method 1 estimates of μ and σ may be used in any problem which requires a long-term estimate of the distribution of fire loss in this sub-population.

5.5 Comparison of results for sprinklered and non-sprinklered buildings

It is not intended to give a detailed comparison of the results obtained for the sprinklered and non-sprinklered sub-populations. This is because the different sub-populations are not homogeneous, there being several factors which vary between sub-populations, particularly between the sprinklered and non-sprinklered sub-populations. Two factors seem particularly important:

(a) Sprinklers are generally installed in those buildings with a greater value at risk. This statement is supported by the fact that from information available on the large fires in the period 1965-72 (Ramachandran[8]) the average gross floor area of sprinklered buildings in the Textile industry was approximately 148,000 sq ft as compared with 60,000 sq ft in the non-sprinklered buildings - a breakdown into single/multi-storey buildings not being available. Assuming, as a first approximation, that monetary value in a building is distributed uniformly, then the value at risk is proportional to the gross floor area.

(b) Sprinklers are generally installed in those buildings which present a greater fire hazard. With an estimated 3,400 sprinklered establishments and only 2,900 non-sprinklered establishments engaged in the Textile industry (see Ramachandran[8]) then using the estimated number of fires given in Table 7, Appendix 1, the expected number of fires per 1000 establishments per annum is approximately 115 for the combined sprinklered sub-populations compared with 100 for the combined non-sprinklered sub-populations, thus supporting the initial statement.

Thus, the differences in the distribution of fire loss between sub-populations is not due entirely to the presence or absence of sprinklers or the building being single- or multi-storey. To make a full assessment of the cost/benefit of sprinklers requires further factors to be brought into the picture. However, the following comments can be made on examination of the results so far obtained:

(i) From Fig.2 and Table 16, Appendix 1, it can be seen that the probability of observing a loss in excess of any given value $x > £100$ is smaller for the sprinklered single-storey and sprinklered multi-storey sub-populations than for the respective non-sprinklered sub-populations.

(ii) From Table 15, Appendix 1, the mean loss in fires in the sprinklered single-storey sub-population is approximately half the mean loss in the non-sprinklered single-storey sub-population. For the multi-storeyed sub-populations the mean loss when sprinklers were present is less than one-sixth of the mean loss with no sprinklers.

(iii) It can be seen from Table 13, Appendix 1 that δ estimated for the sprinklered single-storey and both non-sprinklered sub-populations have almost identical values. Thus, at a given probability level the loss in any one sub-population may be regarded as an approximate multiple of the loss in any other sub-population. More specifically

$$f(x_{S/1}) \simeq f(0{\cdot}50\, x_{NS/1}) \simeq f(0{\cdot}09\, x_{NS/M})$$

in which $x_{S/1}$ is the random variable fire loss in the sprinklered single-storey sub-population, etc.

(iv) From Table 15, Appendix 1, the sprinklered single-storey sub-population accounts for approximately 12 per cent of the fires in the Textile industry but for only 4 per cent of the estimated total loss. These figures for the other sub-populations are approximately 44 per cent and 18 per cent for the sprinklered multi-storey sub-population, 21 per cent and 18 per cent for the non-sprinklered single-storey sub-population, and 22 per cent and 60 per cent for the non-sprinklered multi-storey sub-population. (These percentages do not total 100 due to the effect of the unclassified large losses)

(v) In the sprinklered single-storey, sprinklered multi-storey, non-sprinklered single-storey and non-sprinklered multi-storey sub-populations, the large fire losses account for approximately 60 per cent, 80 per cent, 80 per cent and 90 per cent respectively of the estimated total loss.

Overall the results given in this report show sprinkler installations to have been effective in reducing the total loss due to fire in both single- and multi-storey buildings in use by the Textile industry.

In the Textile industry as a whole, the total loss during the period 1966-72 was estimated at £49.2M at an average of £7.3M per annum or £10,090 per fire at 1966 prices. Large fires accounted for 86 per cent of this total loss, and more than 60 per cent of the large fire loss occurred in non-sprinklered multi-storey buildings. Given a fire which has spread beyond the infant mortality stage, the probability of observing a large fire in the sprinklered multi-storey sub-population is only one-eighth of the same probability for the non-sprinklered multi-storey sub-population. Thus, if sprinklers were to be installed in all multi-storey buildings, the total loss in the Textile industry could be expected to drop considerably. Assuming that all buildings in the Textiles industry have sprinkler installations and using the distribution of fire loss estimated for the two sprinklered sub-populations, then the total loss for the whole Textile industry during the seven-year period 1966-72 would have been approximtely £21M - a reduction of £28M.

The above assumption implies that the sprinklers operated for each fire and that the buildings which were in the non-sprinklered sub-populations are homogeneous with those that were in the sprinklered sub-populations. These assumptions have only a limited validity.

From UK Fire and Loss Statistics[10] over all occupancies where installed, sprinklers failed to operate in approximately 16 per cent of fires. Also, as mentioned earlier, in the Textiles industries sprinklers were, in general, installed in those buildings which presented a greater fire hazard or a greater value at risk. However, the above results show that sprinklers may be expected to reduce the total fire loss to such an extent that mild deviations from the above implied assumptions will not change the overall conclusion that sprinklers are effective in reducing fire loss.

To assess the true worth of sprinklers now requires the above results to be linked with the estimated cost of installation, maintenance, etc of sprinklers. It is intended to make this the subject of a later paper.

Note that the results given in Table 14 are strongly influenced by annual variations in the frequency of fires and of large fires, and by the magnitude of the large losses. The effect of the magnitude of the large losses can only be eliminated by combining the results over several years. Thus, while the annual estimates given in Table 14 show the fire loss situation for a particular year, the real effect of sprinklers and number of storeys in a building on the probability distribution of fire loss can only be fully assessed by considering a period of several years. Therefore, it is imperative that the long-term estimates of μ and σ given in Table 13, Appendix 1, and the long-term estimate of average fire loss given in Table 15, Appendix 1, are used in any further cost/benefit or comparative studies.

6. CONCLUSIONS

(i) For all values of N, and all values of $n \gg 50$, the asymptotic value of Mean ($y_{(m)}$) and Var ($\sigma_{(m)}$) together with the asymptotic result that Cov ($y_{(m)}$, $y_{(l)}$) = Var ($y_{(m)}$), $m > l$ should be used in the estimation procedure.

(ii) For input data which is assumed normally distributed, with up to 20 extremes, and using the asymptotic moments of $y_{(m)}$ the Var-Cov matrix V is well conditioned. The inverse is effectively tri-diagonal. Small changes in the Z -vector are reflected exactly in the estimates of μ and σ .

(iii) It has been shown that estimates based on observations pertaining to only one rank m (= 1, ... R) by the method of moments are subjected to wide between-ranks variations.

(iv) A confidence region based on the assumption that the residuals are normally distributed showed that $\sigma \cos\omega + \mu \sin\omega$ is well determined, and that $\mu \cos\omega - \sigma \sin\omega$ is not well determined, where ω is the angle of rotation. Furthermore, there is dependence between μ and σ . The confidence region is difficult to interpret, and since estimates based on individual extremes will not be used the confidence region may be omitted.

(v) The direct method of using μ_i , σ_i to obtain estimates of the annual total loss is inadequate due to a poor fit in the tail of the distribution. The indirect method where only the unknown part of the distribution is estimated should always be used.

(vi) Where there is sufficient data that 'good' estimates can be obtained based on individual years data, ie Method 2 described in Section 5.1, these estimates should be used to provide an estimate of the annual total fire loss and of the parameters of the parent population.

(vii) Estimates based on several years data averaged within ranks, ie Method 1 described in Section 5.1 should be used in any application that requires an estimate of the parent population over a period of time. If Method 2 fails then Method 1 may be used to obtain an annual estimate of fire loss. However, Method 1 estimates of μ and δ should not be considered as reliable estimates of the parameters of the annual distribution of fire loss.

(viii) Sprinklers have been shown to be effective in reducing the loss due to fire in buildings engaged in the Textile industry.

7. FURTHER WORK

(i) An investigation is required into the effect of heavy-tailed data on the method of estimation. Also to find out if the efficiency of the method may be improved by censoring the upper tail of the data or by systematic sampling of the data so that the extremes used for the estimation cover a greater range of the distribution than the top r extremes alone. This is particularly important for Method 2.

(ii) To apply the procedure given in this report to data from other industries and the distributive trades.

(iii) To make use of the results given in this report in a cost/benefit study of sprinklers.

(iv) It should be noted that the theory and procedure discussed in this report are of general applicability. The computer programs written for this particular study have the lognormal distribution as the assumed parent. However, they are easily adapted to other assumed parent distributions.

(v) At present it is not possible to test for a significant difference between the expected losses estimated for different sub-populations. Much further work is required to develop such a test.

8. ACKNOWLEDGEMENT

Thanks are due to Dr G Ramachandran who gave advice on all aspects of the work discussed in this report.

9. REFERENCES

1. RAMACHANDRAN, G. (1970). Some possible applications of the theory of extreme values for the analysis of fire loss data. Ministry of Technology and Fire Offices' Committee Joint Fire Research Organisation, Fire Research Note 837.

2. RAMACHANDRAN, G. (1972). Extreme value theory and fire losses - further results. Department of the Environment and Fire Offices' Committee Joint Fire Research Organisation, Fire Research Note 910.

3. RAMACHANDRAN, G. (1974). Extreme value theory and large fire losses. ASTIN Bulletin, Vol VII, Pt 3, pp 293-310.

4. RAMACHANDRAN, G. (1975). Extreme order statistics in large samples from exponential type distributions and their application to fire loss. Statistical Distributions in Scientific work, Vol 2, pp 355-367. D Reidel Publishing Company, Dordrecht, Holland.

5. BENCKERT, L G. (1963). The log normal model for the distribution of one claim. ASTIN Bulletin, Vol II, Pt 1, pp 9-23.

6. FERRARA, G. (1971). Distributions des sinistres selon leur coût. ASTIN Bulletin, Vol VI, Pt 1, pp 31-41.

7. LLOYD, E H. (1952). Least squares estimation of location and scale parameters using order statistics. Biometrika, 39, pp 88-95.

8. RAMACHANDRAN, G. (1974). Cost effectiveness of sprinklers. BRE Symposium on Cost Effectiveness and Fire Protection, held at Building Research Station, Garston, September 1974.

9. Private communication (1969).

10. United Kingdom Fire and Loss Statistics, 1972. Department of the Environment and Fire Offices' Committee Joint Fire Research Organisation, HMSO, London 1974. (Annual publication).

11. Standard Industrial Classification, HMSO, London.

12. ROGERS, F E. (1976). Fire losses and the effect of sprinkler protection of buildings in a variety of industries and trades. Forthcoming paper.

APPENDIX 1. TABLES OF RESULTS FOR THE TEXTILE INDUSTRIES

Table 1. Known large losses in sprinklered fires in single-storey buildings in the Textile industry (£'000s)

Rank	1966	1967	1968	1969	1970	1971	1972
1	58	60	60	46	50	223	130
2	40	48	42	20	33	150	110
3	12	39	15	15	23	30	35
4	10	38	11			18	28
5		38	10			16	20
6		23				15	15
7		16				11	
8		10				10	
Total	120	272	138	81	106	473	338

Table 2. Known large losses in sprinklered fires in multi-storey buildings in the Textile industry (£'000s)

Rank	1966	1967	1968	1969	1970	1971	1972
1	90	286	195	100	332	2800	360
2	77	55	91	50	262	1060	100
3	49	53	89	39	115	290	51
4	26	48	38	35	95	170	29
5	25	40	34	25	75	121	20
6	25	30	33	25	33	83	15
7	24	23	29	23	30	62	14
8	17	16	26	20	14	38	13
9	16	10	26	20	10	29	11
10	11	10	25	20	10	20	10
11			21	18		18	
12			18	16		18	
13			13	15		17	
14			10	11		15	
15			10			13	
16			10			10	
Total	360	571	668	417	976	4764	623

Table 3. Known large losses in fires in non-sprinklered single-storey buildings in the Textile industry (£'000s)

Rank	1966	1967	1968	1969	1970	1971	1972
1	143	268	460	400	185	370	125
2	142	109	400	113	58	150	110
3	100	75	320	110	30	110	101
4	65	40	310	93	28	90	80
5	52	40	180	84	25	70	75
6	50	38	140	53	18	50	70
7	25	36	50	49	17	25	55
8	25	32	35	26	15	24	53
9	25	30	27	24	15	22	48
10	20	20	25	21	11	20	40
11	11	18	15	20		18	26
12	10	16	14	20		16	24
13	10	10	13	20			20
14			11	18			16
15				13			15
16				13			14
17				11			12
18				11			10
19				10			
Total	678	732	2000	1109	402	965	894

Table 4. Known large losses in fires in non-sprinklered multi-storey buildings in the Textile industry (£'000s)

Rank	1966	1967	1968	1969	1970	1971	1972
1	400	1033	500	1100	187	800	1502
2	309	300	400	900	175	675	950
3	275	290	315	894	150	508	550
4	257	280	242	864	145	475	450
5	230	203	240	850	63	400	419
6	205	192	205	415	60	350	417
7	172	114	185	379	49	300	400
8	108	112	175	125	45	292	375
9	75	95	155	116	44	271	280
10	71	90	118	100	43	155	275
11	70	85	112	85	41	151	226
12	68	85	68	83	40	150	225
13	58	82	55	75	40	150	190
14	58	64	41	75	38	115	180
15	55	60	40	70	33	110	157
16	48	57	36	62	30	100	139
17	46	50	33	60	30	96	101
18	45	49	33	52	25	90	60
19	41	42	30	50	25	70	51
20	35	40	29	47	25	70	48
21	34	36	28	45	23	65	17
22	25	35	15	38	22	65	17
23	25	34	15	34	21	60	14
24	21	33	13	32	18	55	12
25	18	30	12	32	18	55	11
26	16	29	12	29	15	53	10
27	15	28		20	15	51	
28	14	27		17	14	50	
29	13	23		16	13	48	
30	12	22		13	10	40	
31	12	19		12		36	
32	11	18		12		30	
33	10	18		10		30	
34	10	17		10		22	
35	10	15		10		20	
36		15				18	
37		15				15	
38		12				15	
39		11				15	
40		10				12	
41						11	
42						11	
43						10	
44						10	
Total	2872	3770	3107	6732	1457	6125	7076

Table 5. Unclassified known large losses in the Textile industry (£'000s)

Rank	1966	1967	1968	1969	1970	1971	1972
1	110	34	21	88	315	198	210
2	75	10	10	30	23	150	81
3	18			24	18	105	80
4	16			11	18	70	50
5	15			10	15	45	34
6	10				14	33	22
7					13	30	20
8						13	15
9						10	13
10						10	12
11							11
Total	244	44	31	163	416	664	548

Table 6. Factors used to adjust losses to 1966 price levels

Year	1966	1967	1968	1969	1970	1971	1972
Retail Price Index, base year 1966*	100.0	102.5	107.3	113.1	120.3	131.7	141.0

*Source: Annual Abstract of Statistics 1972, HMSO

Table 7. Estimated parent sample sizes for each sub-population of the Textile industry by year

Sub-population	1966	1967	1968	1969	1970	1971	1972	Avg.
Sprinklered single-storey	80	75	80	86	91	83	82	82
Sprinklered multi-storey	299	280	301	323	342	311	307	309
Non-sprinklered single-storey	139	130	140	150	159	144	143	144
Non-sprinklered multi-storey	151	141	152	163	173	157	155	156

Table 8. Estimates based on observations from one extreme only for the non-sprinklered single-storey sub-population of the Textile industry 1966-72

No. of extreme	μ	σ	$\bar{x}$
1	0.919	0.531	17.487
2	0.059	0.865	8.335
3	-0.624	1.210	11.510
4	-1.047	1.454	24.388
5	-0.667	1.292	17.974
6	-0.816	1.388	25.250
7	-0.295	1.040	8.933
8	-0.102	0.916	7.297
9	0.014	0.854	7.145
10	-0.137	0.927	7.138

Table 9. Annual estimates of μ, σ for sprinklered single-storey sub-population of the Textile industry

		Method 1	Method 2A	Method 2B
1966	μ	-0.616	-0.539	-0.631
	σ	1.024	1.006	1.031
1967	μ	-0.605	-0.529	-0.620
	σ	1.024	1.006	1.031
1968	μ	-0.586	-0.509	-0.601
	σ	1.024	1.006	1.031
1969	μ	-0.562	-0.610	-0.578
	σ	1.024	1.035	1.031
1970	μ	-0.534	-0.583	-0.551
	σ	1.024	1.035	1.031
1971	μ	-0.497	-0.544	-0.512
	σ	1.024	1.035	1.031
1972	μ	-0.467	-0.514	-0.482
	σ	1.024	1.035	1.031

Table 10. Annual estimates of μ, σ for the sprinklered multi-storey sub-population of the Textile industry

		Method 1	Method 2
1966	μ / σ	-1.419 1.340	-1.040 1.124
1967	μ / σ	-1.408 1.340	-1.592 1.421
1968	μ / σ	-1.388 1.340	-0.802 1.107
1969	μ / σ	-1.365 1.340	-0.497 0.892
1970	μ / σ	-1.339 1.340	-2.725 1.949
1971	μ / σ	-1.299 1.340	-2.225 1.958
1972	μ / σ	-1.270 1.340	-1.420 1.294

Table 11. Annual estimates of μ, σ for the non-sprinklered single-storey sub-population of the Textile industry

		Method 1	Method 2
1966	μ / σ	-0.334 1.062	-0.613 1.210
1967	μ / σ	-0.324 1.062	-0.571 1.215
1968	μ / σ	-0.304 1.062	-1.123 1.671
1969	μ / σ	-0.281 1.062	-0.400 1.142
1970	μ / σ	-0.254 1.062	-0.463 0.966
1971	μ / σ	-0.215 1.062	-0.386 1.131
1972	μ / σ	-0.185 1.062	-0.275 1.112

Table 12. Annual estimates of μ, σ for the non-sprinklered multi-storey sub-population of the Textile industry

		Method 1	Method 2
1966	μ / σ	0.401 0.992	0.546 0.889
1967	μ / σ	0.412 0.992	0.648 0.879
1968	μ / σ	0.432 0.992	0.242 1.085
1969	μ / σ	0.455 0.992	0.275 1.171
1970	μ / σ	0.482 0.992	0.592 0.665
1971	μ / σ	0.521 0.992	0.792 0.921
1972	μ / σ	0.551 0.992	0.189 1.326

Table 13. "Best" annual estimates of the parameters of the distribution of fire loss for each sub-population of the Textile industry

	Sprinklered 1-storey		Sprinklered multi-storey		Non-sprinklered 1-storey		Non-sprinklered multi-storey	
	μ	σ	μ	σ	μ	σ	μ	σ
Best long-term estimates at 1966 price level								
	-0.616	1.024	-1.419	1.340	-0.334	1.062	0.401	0.992
Best annual estimates at current price levels								
1966	-0.539	1.006	-1.419	1.340	-0.613	1.210	0.546	0.889
1967	-0.529	1.006	-1.408	1.340	-0.571	1.215	0.648	0.879
1968	-0.509	1.006	-1.388	1.340	-1.123	1.671	0.242	1.085
1969	-0.610	1.035	-1.365	1.340	-0.400	1.142	0.275	1.171
1970	-0.583	1.035	-1.339	1.340	-0.463	0.966	0.592	0.665
1971	-0.544	1.035	-1.299	1.340	-0.386	1.131	0.792	0.921
1972	-0.514	1.035	-1.270	1.340	-0.275	1.112	0.189	1.326

Table 14. "Best" estimates of annual total loss for each sub-population of the Textile industry given at current prices (£'000s)

		SP/1	SP/M	NSP/1	NSP/M	Total un-classified losses	Annual Totals
1966	Sample size	80	299	139	151	6	675
	Known large loss	120	360	678	2872	244	4274
	Total loss	279	655	928	3347		5453
	Average loss	3.5	2.2	6.7	22.2		8.1
1967	Sample size	75	280	130	141	2	628
	Known large loss	272	571	732	3770	44	5389
	Total loss	414	849	971	4189		6467
	Average loss	5.5	3.0	7.5	29.7		10.3
1968	Sample size	80	301	140	152	2	675
	Known large loss	138	668	2000	3107	31	5944
	Total loss	300	968	2178	3520		6997
	Average loss	3.7	3.2	15.6	23.2		10.4
1969	Sample size	86	323	150	163	5	727
	Known large loss	81	417	1109	6732	163	8502
	Total loss	244	752	1414	7129		9702
	Average loss	2.8	2.3	9.4	43.7		13.3
1970	Sample size	91	342	159	173	7	772
	Known large loss	106	976	402	1457	416	3357
	Total loss	284	1343	737	2203		4983
	Average loss	3.1	3.9	4.6	12.7		6.5
1971	Sample size	83	311	144	157	10	705
	Known large loss	473	4764	965	6125	664	12991
	Total loss	630	5102	1276	6569		14241
	Average loss	7.6	16.4	8.9	41.8		20.2
1972	Sample size	82	307	143	155	11	698
	Known large loss	338	623	894	7076	548	9479
	Total loss	502	973	1213	7431		10667
	Average loss	6.1	3.2	8.5	47.9		15.3

Table 15. Total loss in the period 1966-72 for each sub-population of the Textile industry at 1966 price levels (£'000s)

	Sprinklered		Non-sprinklered	
	1-Storey	M-Storey	1-Storey	M-Storey
Sample size	577	2163	1005	1092
Total loss	2110	8731	7597	29107
Average loss	3.7	4.0	7.6	26.7
Sample size	2740		2097	
Total loss	10841		36704	
Average loss	4.0		17.5	
Sample size	4880			
Total loss	49244			
Average loss (Incl. of unclassified large losses)	10.1			

Table 16. Probability that a fire results in a loss greater than x thousand pounds given that the fire has developed beyond the infant mortality stage for each sub-population of the Textile industry

x	SP/1	SP/M	NSP/1	NSP/M
1	0.274	0.145	0.374	0.659
10	0.057	0.035	0.104	0.274
100	0.005	0.005	0.014	0.054

APPENDIX 2

AN ALTERNATIVE METHOD OF ESTIMATION

The method of estimation described here requires further development to obtain estimates which show a clear advantage over estimates obtained by the simpler and quicker Methods 1 and 2. It is included here simply because it provided an alternative to the methods described in Section 5.1 during the development of the technique for estimating the annual distribution of fire loss and annual total loss.

Assume that any composite estimates of μ and σ may be expressed as a linear combination of the annual estimates μ_i, σ_i

ie $$\mu_c = \sum_{i=1}^{N} a_i \mu_i \qquad \sigma_c = \sum_{i=1}^{N} a_i \sigma_i$$

More specifically if μ_c , σ_c are based on L years data ($L \leqslant N$) then $a_i = \frac{1}{L}$ for i' corresponding to a year i which is included in the data leading to the composite estimates, and $a_i = 0$ otherwise.

Now a third method of obtaining estimates of μ_i , σ_i is given by obtaining composite estimates μ_c and σ_c based on different blocks of data chosen in such a way that the equations relating the composite estimates to the annual estimates can be solved simultaneously for μ_i , σ_i . For an N year period composite estimates based on N different blocks of data are required, and must be chosen such that the system of equations has a solution. Therefore denoting the composite estimates obtained from the i^{th} arrangement of the data by μ_{c_i} , σ_{c_i} , then $\underline{\mu}_c = K\underline{\mu}$

where $$\underline{\mu}_c' = \left[\mu_{c_1} \; \mu_{c_2} \; \ldots \; \mu_{c_N}\right]$$

$$\underline{\mu}' = \left[\mu_1 \; \mu_2 \; \ldots \; \mu_N\right]$$

and K is the key matrix, row j of K being the values of a_i , $i = 1, \ldots N$ for the j^{th} arrangement of the data.

Similarly with obvious notation $\qquad \underline{\sigma}_c = K\underline{\sigma}$

$$\therefore \qquad \underline{\mu} = (K'K)^{-1} K' \underline{\mu}_c$$

and $$\underline{\sigma} = (K'K)^{-1} K' \underline{\sigma}_c$$

A solution will exist provided $(K'K)$ is non-singular. The estimates by this method will change with K which should be chosen such that the maximum amount of data is incorporated into the estimation procedure $\lceil K = I$ is equivalent to Method 2$\rfloor$. Of course it is necessary to reduce the data to the same level of prices to apply this method.

This method of estimation was applied extensively to the data for the Textile industry. The composite estimates obtained as the first stage of this estimation method have the same advantages and disadvantages as Method 1 discussed in section 5.2. The estimates of μ_i and σ_i, and the estimates of annual total loss obtained by this method of estimation could not be fully assessed but were often, though not always, in close agreement with those obtained by Method 2. A brief examination of the results of these two methods suggests that they will only be in agreement if the residual variance of each of the estimates $(\mu_{c_i}, \sigma_{c_i})$ $i = 1, \ldots N$ are all approximately equal. If they are not equal then perhaps the matrix $(K'K)^{-1}$ should be adjusted to take into account the variation between the residual variances. This problem has not been looked at in detail, but it is doubtful if any simple adjustment of the matrix $(K'K)^{-1}$ would improve in all cases the annual estimates of μ_i, σ_i by this third method. To improve this method of estimation requires the $Cov(\mu_{c_j}, \mu_{c_k})$, $Cov(\sigma_{c_j}, \sigma_{c_k})$ to be introduced into the solution of the equations giving the annual estimates of μ_i and σ_i. The form of these matrices is not known at present, and their derivation is complex.

Annual estimates provided by this method already require considerably more computer time than either of methods 1 and 2. Further refinements to this method would accentuate this problem still further and although the estimates would be improved it is by no means certain that they would offer any clear advantage over estimates obtained by the very much simpler and quicker Methods 1 and 2. The author does not intend to pursue this method any further.

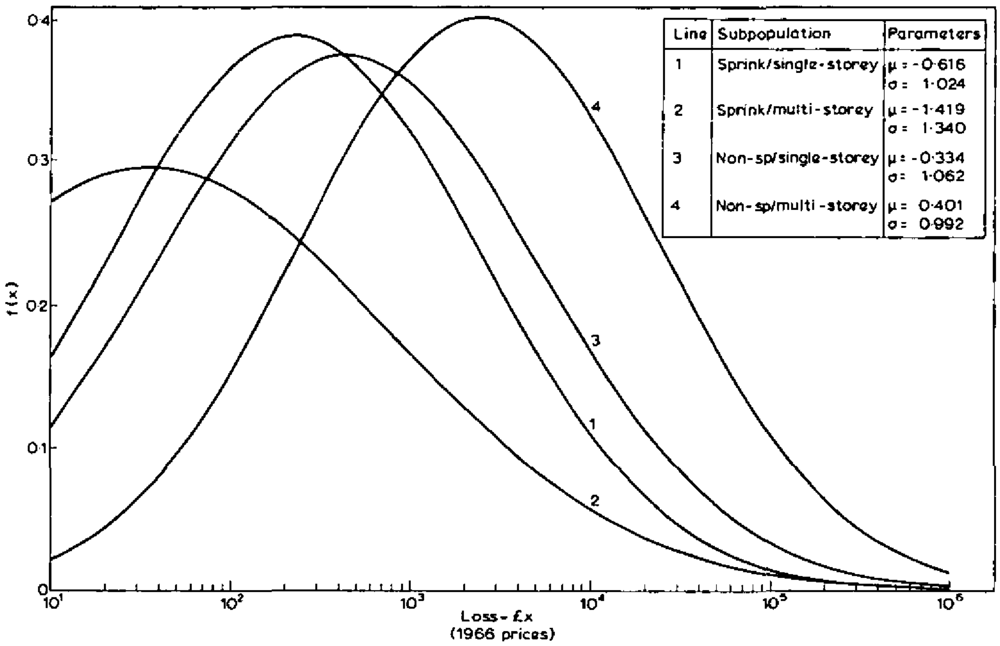


Figure 1 Probability density function of fire loss, for each subpopulation of the Textile industry

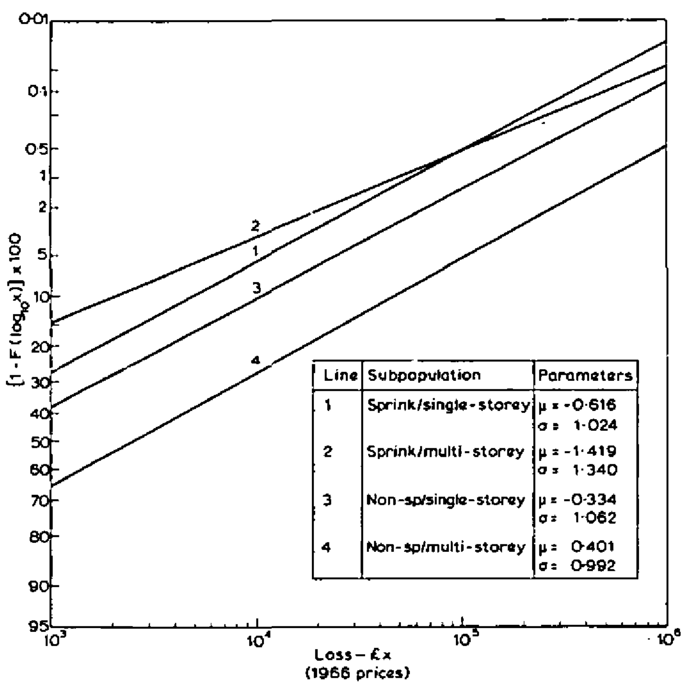


Figure 2 Cumulative distribution function of fire loss, for each subpopulation of the Textile industry