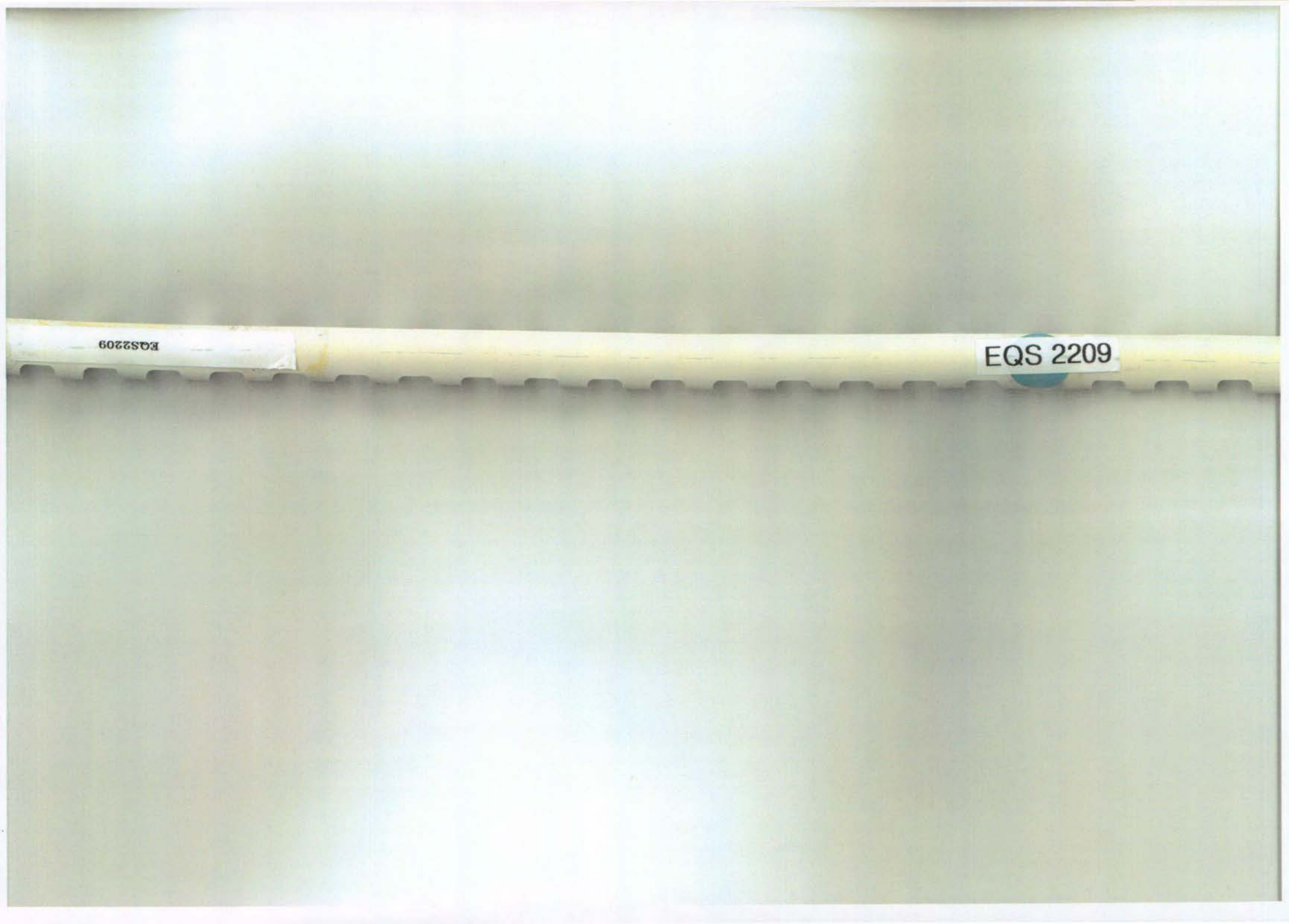
EQS 2209
EQS2209

INDUSTRIAL RESEARCH LIMITED

Improved handling of uncertainties in fault zone hazard estimation

D.A.Rhoades
Applied Mathematics
Industrial Research Limited
P.O.Box 1335, Wellington, New Zealand

R.J. Van Dissen
Institute of Geological and Nuclear Sciences
Lower Hutt, New Zealand

and D.J. Dowrick
Institute of Geological and Nuclear Sciences
Lower Hutt, New Zealand

A report on
research supported by
the Earthquake and War Damage Commission

Improved handling of uncertainties in fault zone hazard estimation

D.A.Rhoades
Applied Mathematics
Industrial Research Limited
P.O.Box 1335, Wellington, New Zealand

R.J. Van Dissen
Institute of Geological and Nuclear Sciences
Lower Hutt, New Zealand

and D.J. Dowrick
Institute of Geological and Nuclear Sciences
Lower Hutt, New Zealand

A report on
research supported by
the Earthquake and War Damage Commission

July 1992

Applied Mathematics
Industrial Research Limited
Wellington
New Zealand

P. O. Box 1335, Wellington, Telephone 0-4-495 5151, Facsimile 0-4-495 5155
situated at 7th Floor, Rankine Brown Building, Victoria University of Wellington

Abstract

Uncertainties in data and parameter values have often been ignored in hazard estimates based on recurrence-time modelling of fault zone rupture. Here, a *mixture of distributions* approach is used to handle uncertainties in parameters estimated from the geological and historical earthquake record of a fault zone, while a *mixture of hazards* approach is used for parameters estimated from a set of similar faults and for data uncertainties. The former approach admits updating of the distributions for uncertainty as time passes, whereas the latter approach does not.

The proposed methods are described in detail for the exponential and lognormal models. A formula for the expected hazard, when the time of the most recent event is uncertain, is derived for the lognormal model. The methods are applied, by way of illustration, to selected faults, namely the Mojave segment of the San Andreas fault, California and the Wellington-Hutt Valley segment of the Wellington fault, New Zealand. The resulting hazard is presented as a single value which takes account of both data and parameter uncertainties, not a range of values.

Contents

1 Introduction

Recurrence-time models are commonly used to estimate the hazard due to the possibility of future large earthquakes occurring in known fault zones (or along known faults). The method involves estimating the recurrence time of faulting either by direct determining of the timing of past earthquakes based on geological studies (i.e. trenching, uplifted Holocene marine terraces) or the historical record, or by estimating the average recurrence interval based on consideration of the single-event displacement size and average slip rate. Combined with a knowledge of the time of the most recent events, recurrence-time models can then give an estimate of the current hazard in the zone. Typically, due to a lack of data on earthquake recurrences on any individual fault or segment, a generic distribution is estimated by combining data from many zones (e.g. Nishenko and Buland, 1987; Jacob, 1984). Sometimes a more elaborate stochastic model is assumed (eg Kiremidjian and Anagnos, 1984), but the data to validate such refinements are even more seriously lacking.

The lack of data makes uncertainties more important, yet methods used to date have tended to ignore uncertainties in many of the estimated quantities, whether geologically based estimates of rupture chronology, or parameters of the statistical distribution. In most cases, they have been geared to point estimates of times of past events inferred from geological data. In practice the uncertainties associated with such estimates are usually large, and may be highly skewed. Davis *et al.* (1989) have shown the importance of taking parameter uncertainties into account in this context. That paper comes closest, in the literature, to the philosophy of the present study.

In some cases where the uncertainties of parameters have been considered the results have been presented as a range of hazard estimates (e.g. Rhoades and Millar, 1983; Brillinger, 1982). These can be helpful in indicating lack of robustness but are often difficult to use in practice,

particularly when the range is wide.

The approach taken here is to try to integrate all important uncertainties into a single estimate for the hazard, rather than produce a range of hazard estimates based on alternative values of uncertain quantities. Uncertainties to be dealt with include those associated with determining the times of occurrence of past earthquakes, the size of single event displacements, the elapsed time since the most recent event, and the parameters of the recurrence-time distribution. The emphasis of this report is on the appropriate statistical methods for estimating hazard in a recurrence-time modelling framework, whether a history of directly determined fault movements, or only the long-term average slip rate and single event displacement is known. More fundamental questions about the appropriateness of recurrence-time modelling in general are not addressed. Primary responsibility for the results rests with Rhoades for the statistical analysis, and with Van Dissen for the assessment of the geological data and their uncertainties.

2 Mathematical formulation of hazard

Seismic hazard may be considered to vary with time t, size of earthquake (often characterised by the magnitude) m, and location z, which may be a point representing hypocentral location or the centroid of the earthquake source, or some discrete zone supposed to enclose the whole source region. Mathematically, hazard may be represented by its conditional intensity (e.g. Rhoades, 1989) or by probabilities. The conditional intensity is a function of time, magnitude and location which is conditional on certain information I (which may include modelling assumptions, parameter estimates and any relevant data). Integrated over a domain of magnitude and location it estimates the instantaneous rate of occurrence of earthquakes within the domain at time t conditional on I. Probabilities can be readily derived from conditional intensities. Let us denote the conditional intensity by $\lambda(t, m, z|I)$ and the probability

an earthquake occurs in the time interval (t_1, t_2), magnitude range (m_1, m_2) and spatial domain Z by $P[E_{(t_1,t_2),(m_1,m_2),Z}]$. Then we have

$$P[E_{(t_1,t_2),(m_1,m_2),Z}|I] = 1 - \exp[-\int_{t_1}^{t_2}\int_{m_1}^{m_2}\int_{Z}\lambda(t,m,z|I)dzdmdt]. \quad (1)$$

In the case of geologically-based estimates of hazard, the location variable is usually dealt with in terms of discrete fault zones, presumed independent. In tracing the history of earthquakes in a single fault zone, it is natural to consider the sequence of the largest events (as measured, in the case of prehistoric earthquakes, by the length of the rupture and the size of the displacement) which have occurred within the zone and to estimate future hazard by extrapolating the statistical properties of this sequence into the future. From the fault displacement history the typical or "characteristic" size of the largest events on the fault can be estimated. If the "characteristic" magnitude is represented by the magnitude range (m_1, m_2) and the fault zone is denoted by Z, then the rate of occurrence $h(t|I)$ of "characteristic" events in the fault zone at time t is

$$h(t|I) = \int_{m_1}^{m_2}\int_{Z}\lambda(t,m,z|I)dzdm. \quad (2)$$

This leads to the following simpler version of equation 1:

$$P[E_{(t_1,t_2)}|I] = 1 - \exp[-\int_{t_1}^{t_2}h(t|I)dt] \quad (3)$$

where $P[E_{(t_1,t_2)}|I]$ is the probability of a "characteristic" earthquake occurring in the fault zone during the time interval (t_1, t_2).

The function $h(t|I)$ is called the hazard function. Suppose that t is measured from the time of the most recent "characteristic" earthquake and that T is the random variable representing the time between the most recent and the next "characteristic" earthquake in the zone. Let $f_T(t|I)$ denote the probability density function of T and $F_T(t|I)$ the cumulative distribution function, i.e.

$$F_T(t|I) = \int_0^t f_T(s|I)ds.$$

Then the hazard function satisfies

$$h(t|I) = \frac{f_T(t|I)}{1 - F_T(t|I)}. \tag{4}$$

The information I on which the hazard estimate is conditioned has, in general, three important components: a model, data, and parameter estimates.

2.1 Model

There are many possible different modelling approaches to hazard in a fault zone, of which only two will be considered here, namely the exponential and lognormal recurrence-time models. The model represents a working hypothesis about the state of nature. One of the purposes in making hazard estimates is to carry out performance tests to confirm or reject particular models (Rhoades and Evison, 1989). The other is for risk assessment and mitigation. In the latter case, it may sometimes be expedient to regard the model as uncertain and to average hazards over several different models. It is not the purpose of this paper to pursue either of these matters in any detail.

2.2 Data

The relevant data typically consist of the dates of historical and prehistoric fault rupture events in the zone, or the average fault slip rate, the displacement associated with a "characteristic earthquake" and the time of the most recent event. These data are usually subject to significant uncertainties, which are nevertheless commonly ignored when it comes to evaluating the hazard. It is a major purpose of this paper to take account of data uncertainties wherever possible. This will be done by describing the data by probability distributions, which are carried through into the estimates of hazard.

2.3 Parameters

The parameters are particular to a model and are estimated from the data. For given data, parameter estimates have a level of uncertainty, sometimes indicated by a variance-covariance associated with point estimates of parameters by standard statistical estimation procedures. When uncertainties in the data are taken into account, the estimated uncertainty in the parameters is increased. Here we shall not be concerned with point estimates at all, but rather with the whole (joint) distribution of parameter values.

2.4 Handling parameter uncertainties

Let θ denote the vector of parameters (associated with a model α) to be estimated from data x. In the simplest (and most unrealistic) case, the data, model and parameter estimates are all regarded as exact and the hazard $h(t)$ is taken to be $h(t|\theta, x, \alpha)$. Probabilities are then estimated by

$$P[E_{(t_1,t_2)}|\theta, x, \alpha] = 1 - \exp[-\int_{t_1}^{t_2} h(t|\theta, x, \alpha)dt]. \tag{5}$$

Bayesian statistical methods (e.g. DeGroot, 1963) take account of uncertainties in parameters by regarding the parameters as random variables. In Bayesian terminology, the *posterior* distribution is the distribution for θ given x (and α). Let the density of this posterior distribution be denoted $f(\theta|x, \alpha)$. Then the probability of a "characteristic" earthquake can be calculated in one of two (non-equivalent) ways:

1. *Mixture of distributions approach*
   Calculate the conditional hazard function $h(t|x, \alpha)$ given by

$$h(t|x, \alpha) = \frac{f(t|x, \alpha)}{1 - F(t|x, \alpha)} \tag{6}$$

   where

$$f(t|x, \alpha) = \int_\theta f(t|\theta, x, \alpha) f(\theta|x, \alpha)d\theta. \tag{7}$$

Then

$$P[E_{(t_1,t_2)}|\boldsymbol{x}, \alpha] = 1 - \exp[-\int_{t_1}^{t_2} h(t|\boldsymbol{x}, \alpha)dt] \tag{8}$$

2. *Mixture of hazards approach*
   Calculate

$$P[E_{(t_1,t_2)}|\boldsymbol{x}, \alpha] = \int_{\theta} P[E_{(t_1,t_2)}|\theta, \boldsymbol{x}, \alpha] f(\theta|\boldsymbol{x}, \alpha)d\theta. \tag{9}$$

   This is equivalent to calculating

$$h(t|\boldsymbol{x}, \alpha) = \int_{\theta} h(t|\theta, \boldsymbol{x}, \alpha) f(\theta|\boldsymbol{x}, \alpha)d\theta. \tag{10}$$

The difference between these two approaches is as follows. In the mixture of distributions approach, the distribution for θ is considered to change as time passes without the earthquake occurring; this is appropriate if θ is estimated from the rupture history of that fault alone. In the mixture of hazards approach, the distribution for θ is considered not to change as time passes without the earthquake occurring; this is appropriate where the distribution for θ is estimated from data across a range of faults considered to have the same recurrence time distribution. In both cases, the non-occurrence of an earthquake on a particular fault is, in principle, new information which may have a bearing on the value of θ, but the distinction is that in the first case the value of θ is considered to be different for different faults, whereas in the second case it is not. Thus, in the first case the impact of the new information is likely to be appreciable; in the second case it is likely to to be negligible.

2.5 Handling data uncertainties

The above formulation assumes both the data and the model to be given. The next step is to take account of the uncertainties in data values, i.e., instead of observing a data vector $\boldsymbol{x}$, we observe its probability density

(or distribution if x is discrete) $f_X(x)$. To take account of the uncertainty in x, we then calculate

$$P[E_{(t_1,t_2)}|\alpha] = \int_x P[E_{(t_1,t_2)}|x,\alpha] f_X(x|\alpha) dx. \tag{11}$$

The mixture of hazards approach is the only option in dealing with the uncertainty of the data because no information on new data can influence the uncertainties associated with past data; past data can only be influenced by improving the techniques used to measure them.

2.6 Handling model uncertainties

Finally, in principle the model itself could be regarded as uncertain. If there are n alternative models $\{\alpha_i, i = 1,\ldots,n\}$ to describe the hazard in a fault zone, where probability $P(\alpha_i)$ has been assigned to model α_i, then we might calculate

$$P[E_{(t_1,t_2)}] = \sum_{i=1}^{n} P[E_{(t_1,t_2)}|\alpha_i] P(\alpha_i) \tag{12}$$

However, it is difficult to see how the probabilities could be assigned, other than subjectively.

3 Recurrence time distribution modelling

Different distributional models for the recurrence time differ in the shape of their hazard functions. The differences are most marked at the extremes of the distribution. The choice of distribution most affects the estimates of hazard immediately after the occurrence of an event and in the upper tail of the distribution, i.e., when the time since the last event is much greater than the mean recurrence time. For times in the middle of the distribution, the choice of model is relatively unimportant. The upper tail probabilities cause a particular quandary because, by definition, there is hardly any data to support an estimate of the shape of the tail.

We now briefly review some common distributional models (Figure 1). The Weibull distribution has been advocated by some authors (e.g. Brillinger, 1982; Sieh *et al.*, 1991) because of its increasing hazard function for values of the shape parameter greater than 1; this shape is considered to be consistent with the fault coming under increasing stress as time goes by without an earthquake occurring. The lognormal distribution has in recent years come into prominence, largely through the work of Nishenko and colleagues (Nishenko, 1985, 1991; Nishenko and Buland, 1987), who have used it to describe a generic distribution for a selection of fault zones on which they have been able to construct fault rupture histories from geological and historical evidence. The initially rising, and then subsiding, hazard function of the lognormal model, is reasonable if one considers that the gradually accumulating stress may be relieved in some other way than by a "characteristic" earthquake, say by a number of small earthquakes, aseismic slip or redistribution of stress within a region. Then again the simplest model of all, the exponential distribution, which has a constant hazard function, has appeal as a model if the timing of events is governed by an essentially chaotic process, where complexity in fault zone strength, redistribution of the stress, and fluctuations in fluid pressure result in effective randomness of events.

It is not the purpose of this paper to lend credence to, test, or favour any of these models. The exponential and lognormal models are considered below, but only as examples of a methodology for dealing with uncertainty in parameters and data.

4 The nature of geological uncertainties

The three most important parameters when using a recurrence-time model (or renewal model) are the elapsed time since the last event, the mean recurrence interval, and the deviation of individual recurrence intervals from the mean. In this section we discuss, using examples

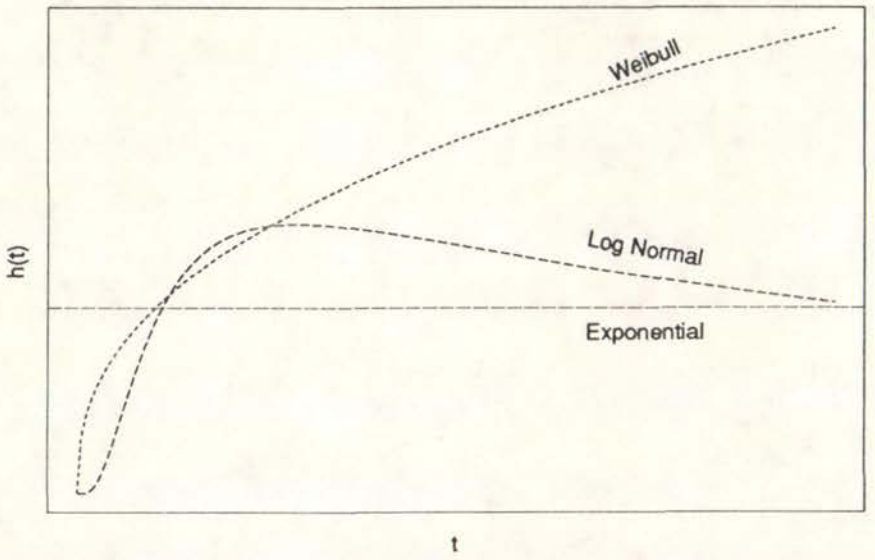

Figure 1: A comparison of Exponential, Lognormal and Weibull hazard functions. All three distributions shown have the same mean; the Lognormal and Weibull distributions have the same coefficient of variation.

from New Zealand and California, how geologists ascertain elapsed time and recurrence interval, and indicate the nature of the uncertainties and assumptions involved. Later we will attempt hazard estimations for some of the faults discussed here.

4.1 Elapsed Time

There are several ways one can hope to determine elapsed time since the last event.

1. *Historical record:* In the Wellington region the historical record is relatively short, and the Wairarapa fault is the only fault for which the historical record can be used to determine elapsed time. 135 years have elapsed since the magnitude 8+ Wairarapa earthquake in 1855.

2. *Constrained by faulted or unfaulted geological deposits of known or inferred age:* For example, Figure 2 is the log of one wall of a trench excavated across the Ohariu fault. Unit 1 is the youngest faulted unit exposed in the trench. Because unit 1 is cut by the fault, it must be older than the most recent surface rupture event. If we knew the age of unit 1, we would know a maximum age for the most recent rupture event (large earthquake) along this portion of the fault. A carbonaceous sample (sample P) from unit 1 was collected, and is dated at 1170 ± 70 yr BP (radiocarbon years before AD 1950). Thus the most recent event along this portion of the Ohariu fault is younger than 1170 ± 70 yr BP, or 960-1160 cal BP (calendar years before AD 1950[1]). This interpretation assumes that sample P was part of unit 1 before unit 1 was faulted, and that sample P is uncontaminated. Given the field relations (not described here), these are not unreasonable assumptions.

   To complete the elapsed time story for this portion of the Ohariu fault, we would like to know the age of the oldest unfaulted deposit. This would constrain the timing of the last event to be younger than 960-1160 cal BP, and older than the as yet undated oldest unfaulted deposit. No dateable material in an "old" unfaulted deposit has been found along this fault; however, it is known from the historical record that this fault has not ruptured within the last 150 years.

---

[1]Carbon-14 dating is one of several dating techniques available for fault studies; a discussion of errors for this technique is given by Pearson and Stuiver (1986) and Stuiver and Pearson (1986). Resolution varies from technique to technique, but commonly the age results are accompanied by some uncertainly term. Lab oriented techniques usually use standard deviation to report uncertainties. Field oriented techniques often handle uncertainty by giving a range of values. The range of values may reflect an "absolute" range, or some sort of "most likely" estimate. The line that separates these two types of estimates is not always clear. For formal incorporation of such data uncertainties into the hazard estimates it is necessary that their meaning be made clear.

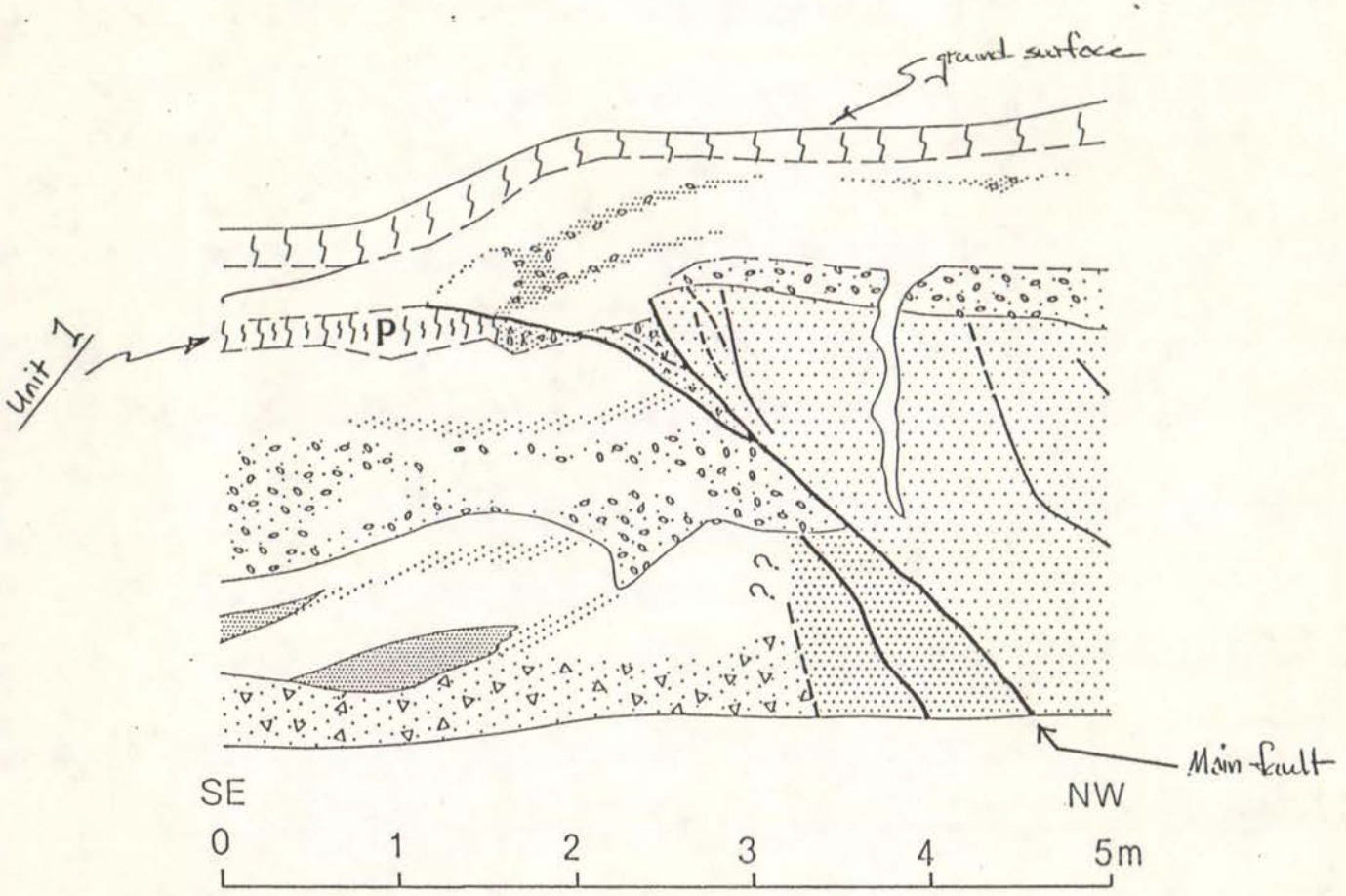

Figure 2: Trench across the Ohariu fault, Wellington region, New Zealand (no vertical exaggeration). Unit 1 is the youngest faulted unit exposed in the trench, and is dated at 1170 ± 70 yr BP (960–1160 cal BP). After Figure 5 of Van Dissen and Berryman (1991).

4.2 Recurrence Interval

The three most common ways to assess earthquake recurrence intervals are using the historical record, geological studies regarding the timing of past events, and inferences based on fault slip rate and size of assumed "characteristic" single event displacement.

1. *Historical record:* Where the historical record is long, and the rates of faulting are high, a rather informative record of the timing of past earthquakes on a specific fault can be obtained from the historical record. In places with a long recorded history there may be a record of several earthquakes rupturing a single fault. In New Zealand, however, where the historical record is relatively short, most active faults have not yet ruptured in a maximum magnitude event, and those that have, have only ruptured once.

2. *Paleoseismicity studies:* Geological studies are increasingly being used to extend the earthquake history on a fault beyond the historical, and have added greatly to our knowledge of the periodicity and size of past fault movements. A primary aim of these studies is to directly determine the timing of past rupture events (assumed to represent past large earthquakes). Two examples are given below.

   For example, Figure 3 is the log of a trench excavation across the Wellington fault near the south coast. Unit 4 is the youngest unit exposed in the trench that is deformed by faulting. Unit 4 is thus older than the faulting event, and a wood sample taken from unit 4 has an age of 790-930 cal BP. Unit 5 is undeformed by the faulting that bounds unit 4, thus it must be younger than the faulting event. A wood sample from unit 5 has an age of 560- 670 cal BP. These relationships demonstrate that surface rupture along this section of the Wellington fault occurred sometime between the two constraining dates of 560-670 cal BP and 790-930 cal BP; that is, rupture occurred sometime between 560 and 930

cal BP. From evidence elsewhere along the fault, we can further constrain the timing of this event to 670-830 cal BP. We have also identified a younger event at 300-450 cal BP. If we knew the timing of several other older events we would be able to calculate an average recurrence interval based on geologically determined timings of past events.

One of the classic paleoseismicity studies comes from Pallett Creek along the Mojave segment of the San Andreas fault. Using geological methods (trenching), and high- resolution carbon-14 dating, workers have been able to constrain the timing of the past dozen or so earthquakes (Figure 4). From these data, Sieh *et al.* calculate an average recurrence interval of faulting of about 132 years. This recurrence interval assumes that no earthquakes have been missed.

3. *Inferences based on slip rate and size of single event fault displacement:* Ideally, in order to determine average recurrence interval for a particular section of fault, one would like to know the timing of the most recent earthquakes along that section of fault. Unfortunately, the historical record in New Zealand is relatively short, and productive trenching sites are not easy to find. Along the Wellington-Hutt Valley segment of the Wellington fault about half a dozen trenches have been excavated and about a dozen radiocarbon dates obtained, but still there are constraints on the timing of only the two most recent earthquakes (Van Dissen *et al.*, 1992). In lieu of a complete history of recent earthquake timing, a general estimate of the average earthquake recurrence interval for a given fault can be inferred by dividing what is considered to be the "characteristic" single event displacement size by the fault's average slip rate. An example from the Wellington fault, using data from Berryman (1990) is given below.

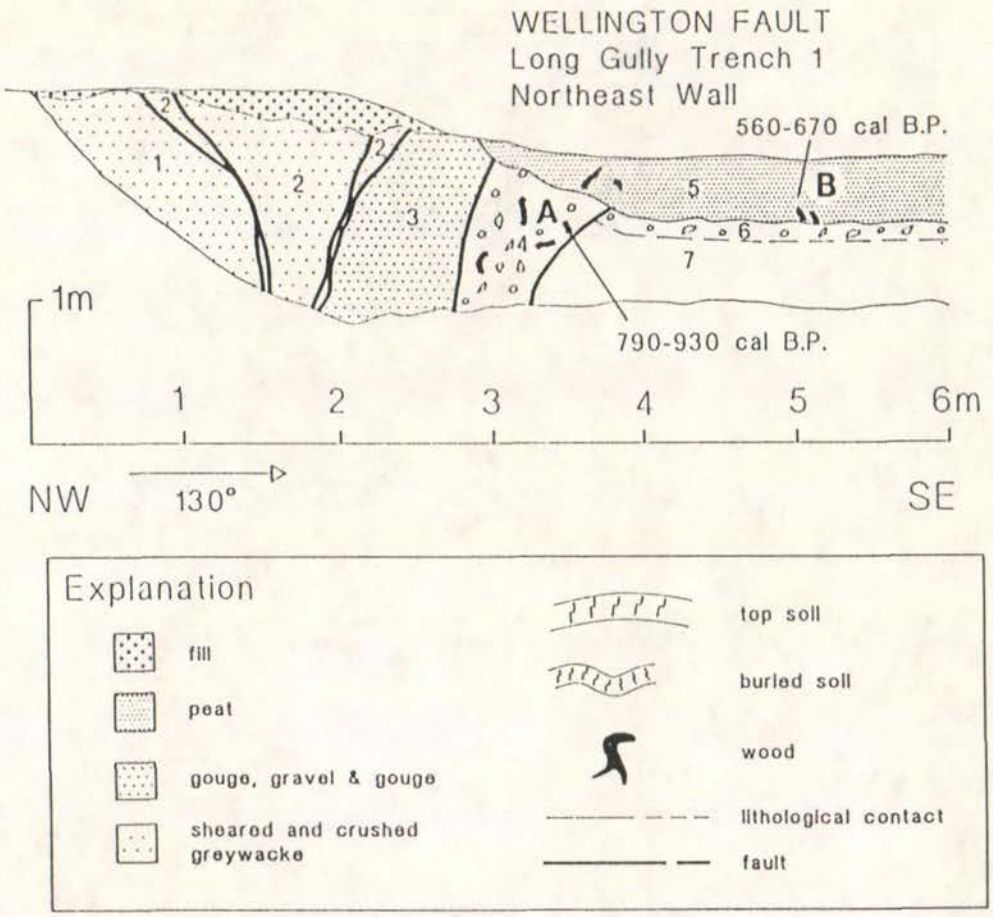

Figure 3: Trench across the Wellington fault. From Figure 4 of Van Dissen *et al.* (1992).

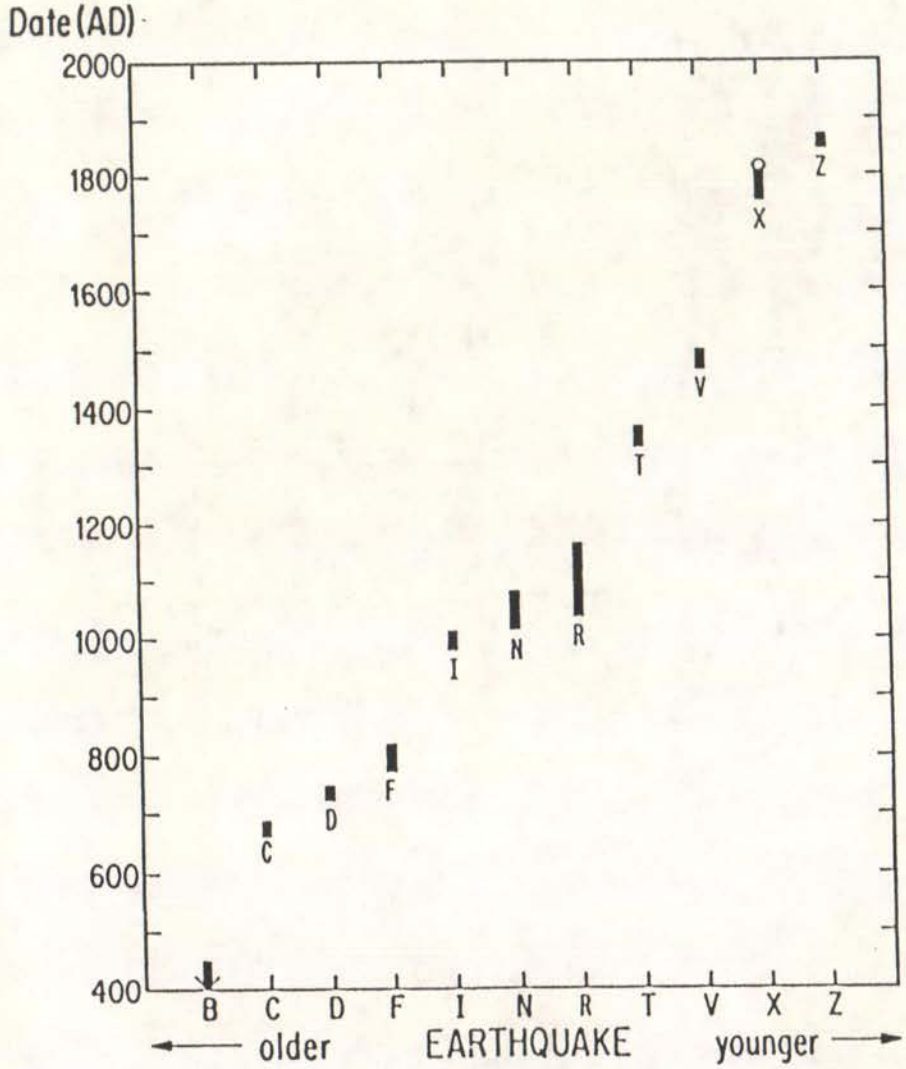

Figure 4: Estimates of the dates for earthquakes recorded at Pallett Creek. Bars give 95% confidence intervals. From Figure 6 of Sieh *et al.* (1989).

A faulted Holocene river terrace sequence at Te Marua (Figure 5), near Upper Hutt, records lateral offsets associated with the last five movements along the southern portion of the Wellington fault. The first terrace above river level (T1) is not displaced and must have formed after the latest faulting event. Two channels on the next highest terrace (T2) are right-laterally offset by 3.7 and 4.7 m (I-I′ and J-J′). The riser to the next highest terrace (R2) is laterally offset by 7.4 m (K-K′). The 3.7 and 4.7 m offsets represent the most recent single-event displacement; the 7.4 m offset probably represents a two-event offset. Under the assumption that similar displacements occur in successive earthquakes, the c. 11 m difference between the R3 or R4 offset (c. 18.5 m, L-L′ and M-M′), and the R2 offset (7.4 m) probably represents three individual fault movements. This interpretation would indicate that single-event horizontal displacements at Te Marua range from about 3.2–4.7 m.

The Wellington fault at Emerald Hill, also near Upper Hutt, is interpreted to have a relatively constant average lateral slip rate of 6.0–7.6 mm/yr (Figure 6), based on three faulted terraces that are laterally displaced by 104 ± 10, 437 ± 20, and 940 ± 40 m, and have assigned ages of 14 ± 4 ka, 70 ± 5 ka, and 140 ± 10 ka respectively (where 1 ka = 1,000 years). The uncertainties associated with the displacements reflect the difficulties in measuring the offsets in the field. Uncertainties can arise from difficulties in correlating offset features across the fault, and projecting offset features into the fault. The terrace ages were assigned based on loess and tephra stratigraphy and correlations with climatic events of "known" age. The terraces were not directly dated. The uncertainty associated with their age reflects how well the age of a given climatic event can be constrained, and does not reflect the possibility that the terraces were miss-correlated.

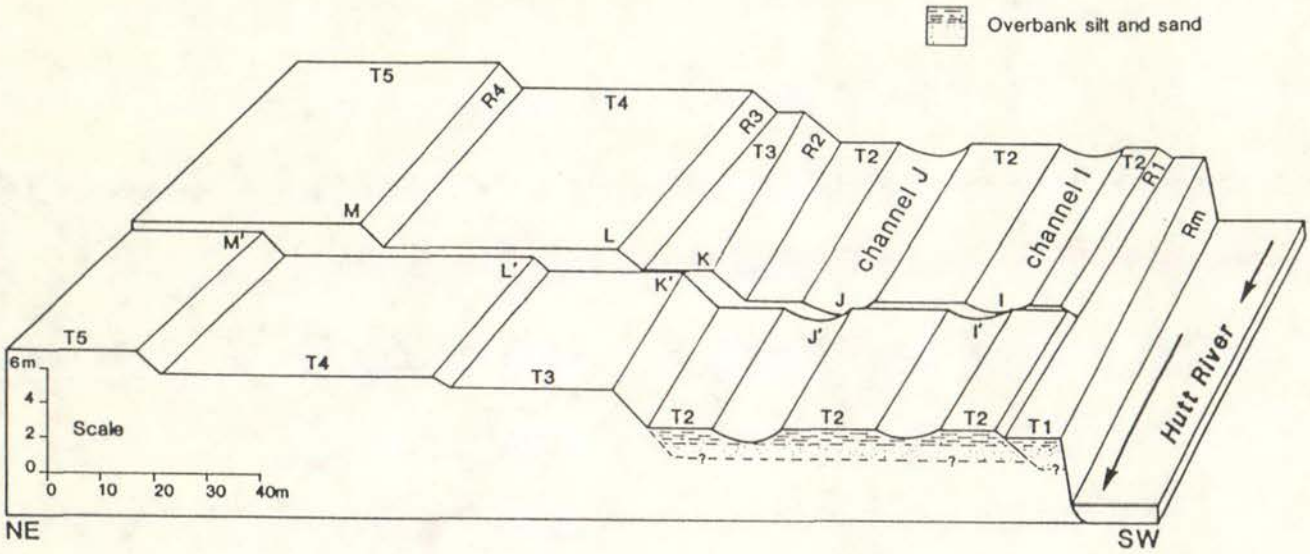

Figure 5: Holocene fluvial channels and terraces displaced by the Wellington fault at Te Marua. Offsets I-I′, J-J′, K-K′, L-L′ and M-M′ are 3.7, 4.7, 7.4, 18.0 and 19.0 m, respectively. After Figure 11 of Van Dissen *et al.* (1992).

Because there is no evidence that the fault is creeping, it is assumed that offsets are the result of coseismic (earthquake) displacements. By dividing the Te Marua single- event displacement of 3.2-4.7 m by the Emerald Hill slip rate of 6.0-7.6 mm/yr a fault-rupture recurrence interval of 420- 780 years is calculated. The 360 year range of this interval reflects uncertainties associated with both size of single event displacement, and slip rate. This range does not provide any direct measure of the variability of individual recurrence intervals about an average value; there is no way of directly determining the coefficient of variation of the recurrence time distribution.

The above illustrates the kind of assumptions that must be made in order to use single event displacement size and slip rate to estimate recurrence interval. In the Wellington case there is an assumption that the slip rate calculated at Emerald Hill, based on offsets ranging in age from 14 ka to 140 ka, also applies to the faulted terraces at Te Marua that are all younger than several thousand years. The close proximity of these two sites (within 1.5 km from each other), the constant strike and simple trace of the fault in this area, and the observation that the slip rate at Emerald Hill has been constant on the scale of several tens of thousands of years (Figure 6) give some support to the assumption that the slip rate has been constant over the last several thousand years. There is also the assumption that earthquakes on this fault have a "characteristic" size of 3.2–4.7 m as measured at Te Marua for the past five earthquakes. Such assumptions are bound to be regarded as reasonable by some and as unreasonable by others.

Several general points emerge from the preceeding discussion and examples regarding the nature of geological uncertainties.

- Not all uncertainties in paleoseismicity data can be expressed as a standard deviation; some can only be expressed as a range with some uncertainty about the endpoints. The sampling methods

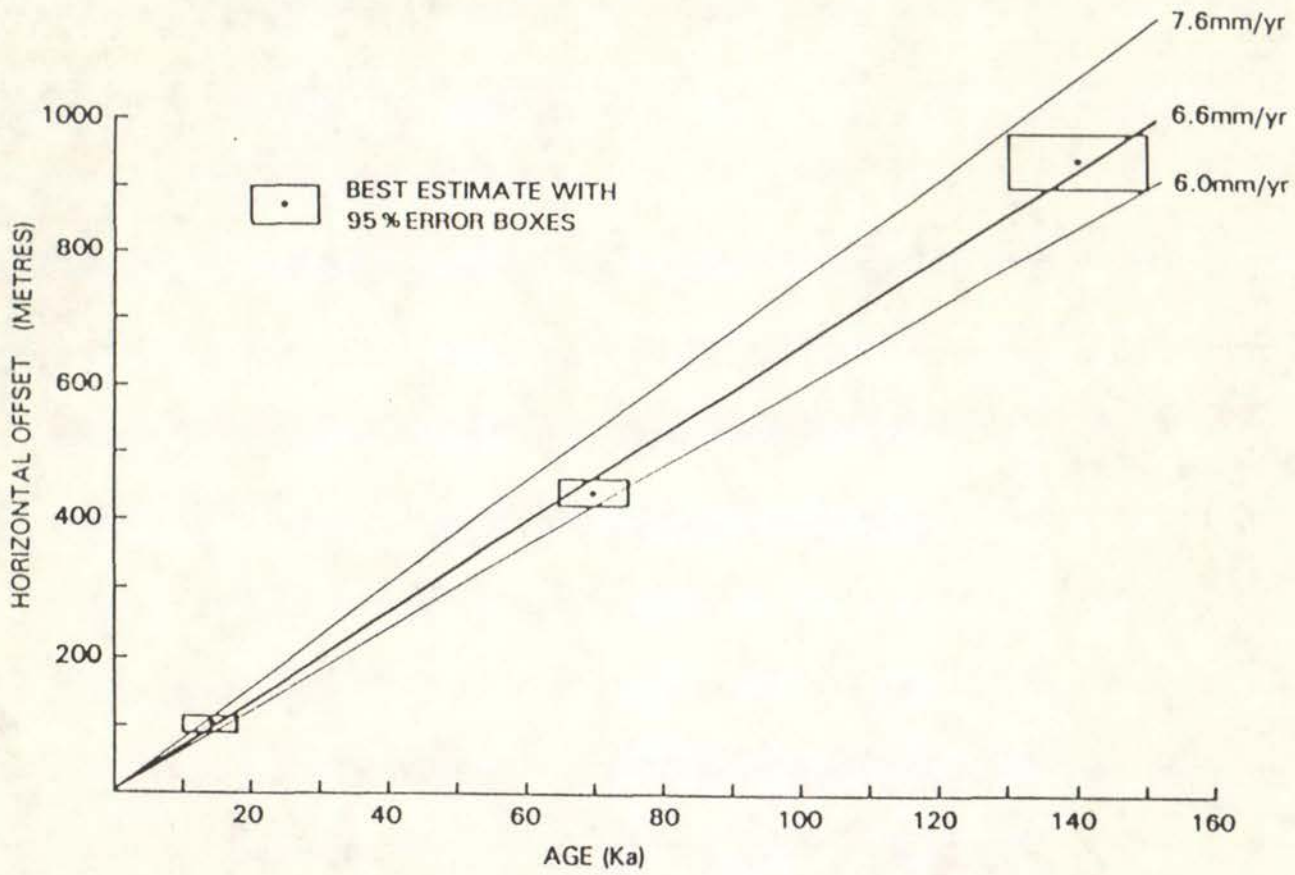

Figure 6: Slip rate on the Wellington fault at Emerald Hill. The points represent the best estimates of offsets and ages. Boxes approximate 95% confidence limits. From Figure 7 of Berryman (1990).

used below for handling data uncertainties can easily cope with the various cases that arise in practice.

- Earthquake history data is incomplete for the vast majority of faults in New Zealand. Recurrence interval data obtained from directly determining the timing of past rupture events is rare, and difficult to obtain. In applying models other than the simplest (i.e. exponential) it is necessary to assume a value or distribution for, say, the coefficient of variation and to be prepared to investigate the sensitivity of the hazard to the value of this parameter.

5 Uncertainties in the exponential/Poisson model

If the recurrence-times in a fault zone are independent and identically distributed exponential random variables with mean $1/\mu$, then the sequence of rupture events is a Poisson process with rate parameter μ. The recurrence-time distribution has density

$$f(t|\mu) = \mu \exp(-\mu t) \tag{13}$$

and constant hazard function

$$h(t|\mu) = \mu \tag{14}$$

If the available data x consist of a sequence of past recurrence intervals, $T_i, i = 1, \ldots, k$, then, assuming a non-informative prior for μ, the posterior density for μ is proportional to the likelihood function, i.e.,

$$f(\mu|T_1, \ldots, T_k) \propto \mu^k \exp(-\mu \sum_{i=1}^{k} T_i) \tag{15}$$

This density is independent of the individual intervals; it depends only on the total time and the number of events. Thus the same density applies if the total time and number of events are the only available data.

In the latter case the number of events is, in practice, uncertain also, commonly being estimated from the total displacement of "known" age and an estimated displacement in a characteristic event. There are at least two ways in which the data errors could be handled, depending on the way the data are deduced from the geological information.

1. A discrete distribution could be estimated subjectively by the geologist for the number of events over the time period and a distribution for the time period uncertainty determined from the dating procedures.

2. A distribution for the size of the "characteristic" earthquake could be estimated (again subjectively), and a distribution for the total displacement over a given time period. These distributions determine the distribution for the number of events over the period.

6 Uncertainties in the lognormal model

If the recurrence times in a zone are independent and identically distributed lognormal random variables with mean γ and coefficient of variation δ, then the sequence of events may be either more regular (most values close to the mean) or more clustered (most values much less than or much greater than the mean) than a Poisson process according to whether $\delta < 1$ or $\delta > 1$. Thus the lognormal family can represent a wide range of distribution shapes (Figure 7) as can other families such as Weibull and Gamma. On the other hand, the tail shape of the lognormal, and hence the hazard at times much longer than the mean recurrence interval, is particular to that distribution, as discussed above.

The lognormal density is given by

$$f(t|\mu,\sigma) = [t\sqrt{2\pi}\sigma]^{-1} \exp[-\frac{1}{2}\frac{(\log t - \mu)^2}{\sigma^2}] \qquad (16)$$

where μ and σ are the mean and standard deviation of the (normal) distribution for the logarithm of recurrence intervals. An alternative

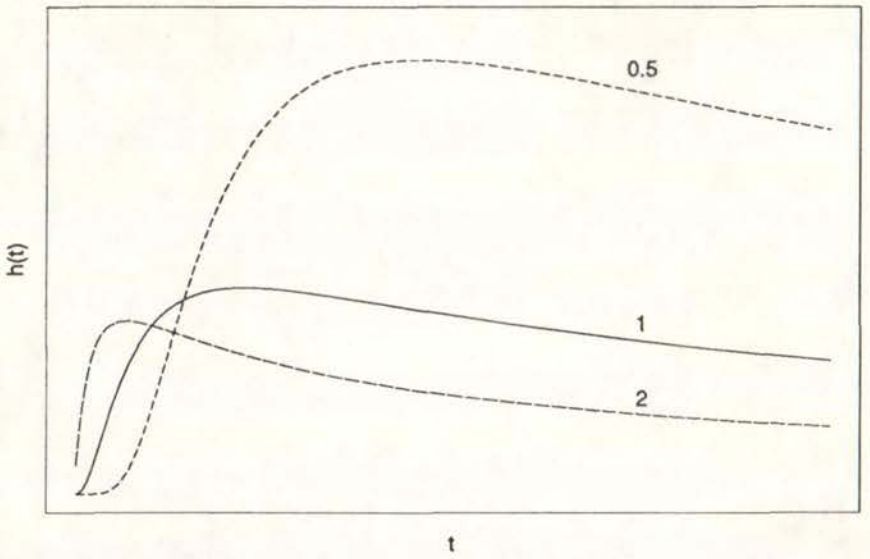

Figure 7: Log normal hazard funtion for different values of the coefficient of variation, δ=0.5, 1 and 2.

parameterisation is in terms of the mean γ and coefficient of variation δ, where $\gamma = \exp(\mu + \frac{1}{2}\sigma^2)$ and $\delta = \sqrt{\exp(\sigma^2) - 1}$. Maximum likelihood estimation of μ and σ (for the normal distribution), by the sample mean and standard deviation, is equivalent to maximum likelihood estimation of γ and δ (Johnson and Kotz, 1970, p119).

6.1 Case where past events on fault are dated

Let us consider the problem of estimating the density $f(\mu, \sigma|T_1, \ldots, T_k)$. If, in the Bayesian formulation, we choose a non-informative prior (i.e. a prior distribution which weights all values of the parameter equally), then the posterior density is proportional to the likelihood function, i.e.,

$$f(\mu, \sigma|T_1, \ldots, T_k) \propto \sigma^{-n} \prod_{i=1}^{n} \exp[-\frac{1}{2}\frac{(\log T_i - \mu)^2}{\sigma^2}]. \qquad (17)$$

Davis *et al.* (1989) incorporated the time since the last earthquake (the *drought*) into the likelihood function. An alternative approach,

adopted here, is to use equations 6 and 7 with the density $f(\theta|x;\alpha)$ given by equation 17; the additional information on the drought is then incorporated automatically by the hazard function, which changes the weighting of the possible values of θ as time passes without an earthquake occurring.

Davis *et al.* (1989) also gave a method for incorporating the data uncertainty into the hazard estimate, but their method did not allow for dependence which arises because the interevent times are linked across a common boundary. The method adopted here allows for this dependence by sampling from the distribution of uncertainties in each independent dating interval. In some cases the date of a particular event may be represented by, say, a normal distribution centred on a best estimate. In other cases the date of an event may only be known to lie between two such dates, each having its own uncertainty. Because of the diversity of situations that may arise, the proposed sampling procedure for handling data uncertainties is considered more satisfactory than using an oversimplified model which can be handled analytically.

6.2 Case where past events on fault are not dated

If past fault movements are not individually dated, i.e, if only the average slip rate and size of "characteristic" event can be estimated from geological evidence, then one is in the position of having to use a generic distribution, estimated, say, from faults in a similar tectonic setting. Only the mean of this distribution can be estimated from data pertaining to the fault zone at hand; the coefficient of variation would have to be obtained from the generic distribution. The selection of similar fault zones for estimating the generic distribution is a matter for expert geological judgement, not dealt with here. The spread parameter σ (or, equivalently, the coefficient of variation δ) is now estimated independently of the fault zone at hand, so its distribution is not open to adjustment as time passes without an earthquake occurring

in the zone. However, the distribution of μ (which, together with σ or δ, determines the mean γ) is open for adjustment, since it is estimated using data from the fault zone. The procedure is thus different from the case above. Now we have

$$h(t) = \int_x \int_\sigma h(t|\sigma, x) f(\sigma, x) d\sigma dx \tag{18}$$

where

$$f(t|\sigma, x) = \int_\mu f(t|\mu, \sigma, x) f(\mu|\sigma, x) d\mu \tag{19}$$

In other words, the usual mixture of distributions approach is used for μ, and the mixture of hazards approch for σ. There are several possible approaches to obtaining an approximate likelihood function for μ. Two approaches are given here, but others are possible.

1. *Large k: normal approximation*
   If the total interval $\sum_{i=1}^{k} T_i$ for a large number of displacements k has been estimated, then the central limit theorem ensures that $(\overline{T} - \gamma)/Var(\overline{T})$ has approximately a standard normal distribution (where $\overline{T} = \sum T_i/k$) and hence

$$f(\gamma|\delta, \sum T_i) \propto \frac{1}{\gamma} \exp[-\frac{1}{2k}(\frac{\sum T_i - \gamma}{\gamma\delta})^2]. \tag{20}$$

2. *Small k: an ad hoc approach*
   For the case where the number of displace ments k is small, the question of how to estimate $f(\mu|\sigma, x)$ is answered here (somewhat tentatively) in a rather *ad hoc* and pragmatic way. It is possible that a theoretically more consistent approach can be developed. Ideally, if we choose a non-informative prior, the posterior distribution would be proportional to the lognormal likelihood for μ, based on the total of k recurrence intervals. However, this is not very tractable as the lognormal likelihood depends on the unmeasured quantity $\sum_{i=1}^{k} \log T_i$ rather than $\sum_{i=1}^{k} T_i$. Instead, it is proposed to

approximate the lognormal likelihood by a gamma likelihood, the gamma distribution having the convenient reproducing property that the sum of independent gamma random random variables is a gamma random variable. The gamma density, like the lognormal, is capable of assuming a variety of shapes depending on parameter values. Forcing the mean and coefficient of variation to be the same as the desired lognormal distribution should ensure that the shape of the distribution used is not too different from the desired distribution (Figure 8).

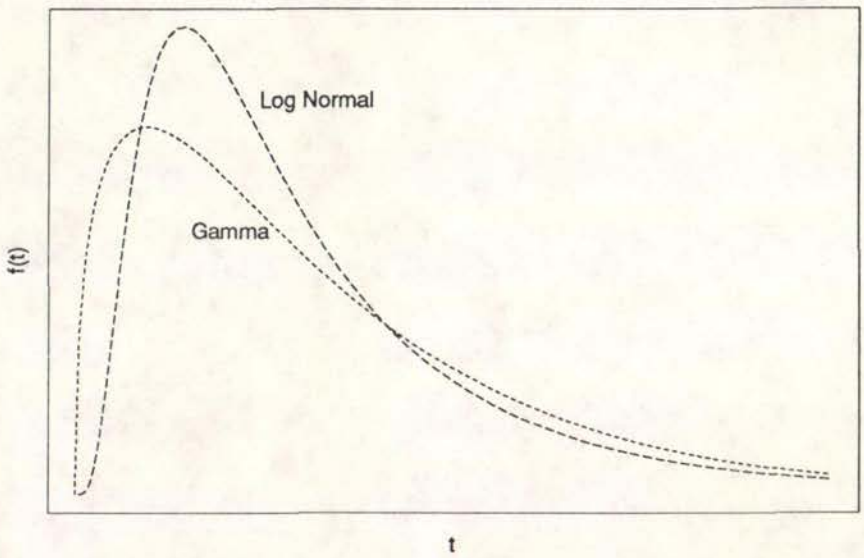

Figure 8: A comparison of density functions of Gamma and Lognormal distributions with the same mean and coefficient of variation.

The gamma density can be written in terms of two parameters α and λ, thus:

$$f(t|\alpha,\lambda) = \frac{\lambda^{\alpha}}{\Gamma(\alpha)} t^{\alpha-1} \exp(-\lambda t). \tag{21}$$

This distribution has mean α/λ and coefficient of variation $1/\sqrt{\alpha}$. The sum of k independent Gamma(α,λ) random variables is a Gamma$(k\alpha,\lambda)$ random variable. Hence, if the time over which k

fault movements have occurred is $\sum_{i=1}^{k} T_i$, and the coefficient of variation δ has been estimated from external information, then α is estimated by $1/\delta^2$ and the density for λ (assuming a non-informative prior) satisfies

$$f(\lambda|\alpha, \sum_{i=1}^{k} T_i) \propto \lambda^{k\alpha} \exp(-\lambda \sum_{i=1}^{k} T_i). \tag{22}$$

Changing variables, the density for the mean γ $(= \alpha/\lambda)$ of the recurrence time distribution then satisfies

$$f(\gamma|\alpha, \sum_{i=1}^{k} T_i) \propto \gamma^{-k\alpha-2} \exp(-\alpha \sum_{i=1}^{k} T_i/\gamma). \tag{23}$$

Changing variables again to the lognormal parameters μ and σ, we have

$$f(\mu|\sigma, \sum_{i=1}^{k} T_i) \propto \gamma(\mu,\sigma)^{-k\alpha-1} \exp[-\frac{\sum_{i=1}^{k} T_i}{\delta^2(\sigma)\gamma(\mu,\sigma)}] \tag{24}$$

where $\gamma(\mu,\sigma) = \exp[\mu + \frac{1}{2}\sigma^2]$ and $\delta(\sigma) = \sqrt{\exp(\sigma^2) - 1}$.

6.3 Case where the time of the last movement is only known vaguely

Sometimes it may not be known when the last movement occurred on a fault. If there is no information on the time of the last movement then the current hazard under the lognormal model is the same as under the exponential model (i.e., it depends only on the mean recurrence time). However, if it is known that the time since the last movement is between a and b (where b may be ∞), then the following result may be used to compute the expected hazard.

For lognormal recurrence times, the expected hazard given that the time since the last movement is between a and b is given by

$$E[h(t)|a < t < b] = \frac{\Phi(\frac{\log b-\mu}{\sigma}) - \Phi(\frac{\log a-\mu}{\sigma})}{b[1-\Phi(\frac{\log b-\mu}{\sigma})] - a[1-\Phi(\frac{\log a-\mu}{\sigma})] + \gamma[\Phi(\frac{\log b-\mu-\sigma^2}{\sigma}) - \Phi(\frac{\log a-\mu-\sigma^2}{\sigma})]} \tag{25}$$

where Φ is the standard normal probability integral and γ is the mean of the distribution.

This result is proved in the Appendix.

In order to exploit the above result, which does not admit mixing of distributions, it is necessary to use the method of hazards approach for both parameters. When the time of the last event is not known with any precision, the amount of new information (about μ) derived from the persistence of the drought is relatively small. Hence it may be acceptable to use the mixture of hazards approach for μ as well as for σ. Thus

$$P[E_{(t_1,t_2)}] = \int_x \int_\sigma \int_\mu P[E_{(t_1,t_2)}|\mu, \sigma, x] f(\mu, \sigma, x) d\mu d\sigma dx. \tag{26}$$

Where it is considered unacceptable to use the mixture of hazards approach for both parameters, the uncertainty in the elapsed time since the last fault rupture can be handled by sampling, as for other data.

7 A description of numerical sampling of distributions

The estimation of hazard under the general approach outlined above formally involves the evaluation of multiple integrals which could occasionally be accomplished analytically but more often by resort to numerical integration. In order to give maximum flexibility in coping with a variety of special cases which may arise in practice, it is convenient to carry out the the integration approximately by taking averages over samples generated to conform to the the relevant distributions for the data and paramaters. The precision of the integration depends on the size of the samples chosen. First, a sample $x_1, \ldots, x_n$ is drawn from the distribution of data with density $f(x)$. For each x_i a sample $\theta_1, \ldots \theta_k$ is drawn from the conditional distribution of parameters with

density $f(\theta|x_i)$. The density $f(t|x_i)$ is estimated (c.f. equation 7) by

$$\overline{f(t|x_j)} = \frac{\sum_{i=1}^{k} f(t|\theta_i, x_j)}{k}. \tag{27}$$

Then, using equation 12 we estimate $P[E_{(t_1,t_2)}]$ by

$$\overline{P[E_{(t_1,t_2)}]} = \frac{\sum_{j=1}^{n} P[E_{(t_1,t_2)}|x_j]}{n}. \tag{28}$$

Equivalently, the hazard function $h(t)$ is estimated by

$$\overline{h(t|\alpha)} = \frac{\sum_{j=1}^{n} h(t|x_j)}{n}. \tag{29}$$

The method used to generate the samples in the examples below is now described.

7.1 Generating random samples from a distribution

Let f be a probability density for a vector random variable $Y = (Y_1, \ldots, Y_n)$ (or a function, such as a likelihood function, which is proportional to this density). We wish to generate a pseudo random sample of size m conforming to the density f. This may be done as follows.

1. For each Y_i find an interval (a_i, b_i) such that f takes on negligibly small values when Y_i is outside (a_i, b_i).
2. Find a number f_{max} which is $\geq$ the maximum value of f.
3. Generate a uniform pseudo-random vector y in the n-dimensional box defined by $\{(a_i, b_i), i = 1, \ldots, n\}$.
4. Evaluate $f(y)$.
5. Generate a uniform(0,1) pseudo random number u.
6. Include y in the sample if $u < f(y)/f_{max}$, otherwise not.
7. Repeat steps 3-6 until a sample of size m has been obtained.

In the case of parameter uncertainties, other approaches could have been adopted. For example, numerical integration might have been attempted, or a weighting of each sample of parameter values by the likelihood function. However the above sample selection procedure leads to the more straightforward numerical procedures, which only involve simple averaging. It also allows a similar approach to be used for generation of samples from data and parameter distributions.

7.2 Summary of steps for the hazard estimation

1. Obtain sample $x_1, \ldots, x_n$ from data distribution $f(x)$.
2. For each $x_j, j = 1, \ldots, n$
   - Obtain sample $\theta_1, \ldots, \theta_k$ from parameter distribution $f(\theta|x_j)$
   - Perform distribution mixing
   $$f(t|x_j) = \frac{\sum_i f(t|\theta_i; x_j)}{k}$$
   - Calculate hazard function
   $$h(t|x_j) = \frac{f(t|x_j)}{1 - F(t|x_j)}$$
3. Perform hazard mixing
$$h(t) = \frac{\sum_j h(t|x_j)}{k}$$
4. Calculate probability of earthquake in time period of interest
$$P[E_{(t_1,t_2)}] = 1 - \exp[-\int_{t_1}^{t_2} h(t)dt]$$

The details of the procedure differ depending on the nature of the data but the general procedure is always the same. The only qualification is that in some cases a parameter value estimated externally from the data of the fault zone being studied is included in the "data", not in the "parameters".

8 Examples

By way of illustration, the above methods are applied to two examples: the Mojave segment of the San Andreas fault, California, where individual events at Pallett Creek have been dated (Sieh *et al.* 1989); and data from the Wellington–Hutt Valley segment of the Wellington fault where, except for the last two displacements, the individual events have not been dated (Van Dissen *et al* 1992). Both of these fault segments were discussed in Section 4 above.

8.1 Pallett Creek

The past dozen or so movements on the San Andreas fault at Pallett Creek have been dated by Sieh *et al.* (1989), as in Table 1. Apart from the two most recent movements, the dates of which are known from historical records, the dates are determined from geological information and are given by Sieh *et al.* (1989) as middle estimates and 95% confidence intervals. They are interpreted here by us as independent normal estimates $\pm$ 2 standard deviations. These uncertainties may not be the uncertainties which we would use if we were interpreting the geological evidence ourselves; in particular it is doubtful that the dates can realistically be considered as independent in all cases. However, for the purposes of illustration the dates are regarded as independent and normally distributed with standard deviation as given in the last column of Table 1.

Table 1
Estimated dates of occurrence for events at Pallett Creek
(from Table 3 of Sieh *et al.*, 1989).

Event	Date (A.D.)	95% confidence interval	std dev'n
Z	1857. 1.9	NA	0
X	1812.12.8	NA	0
V	1480	(1465-1495)	7.5
T	1346	(1329-1363)	8.5
R	1100	(1035-1165)	32.5
N	1048	(1015-1081)	16.5
I	997	(981-1013)	8
F	797	(775-819)	11
D	734	(721-747)	6.5
C	671	(658-684)	6.5
B	before 529	NA	NA

For sampling from the distributions of data and parameters, an arbitrary sample size of 50 has been chosen. Although much smaller than the sample sizes used in some simulations, this seems large enough to ensure robustness in the results for the present examples.

Fifty simulations of date sequences for events C to Z were made, the date for each event being generated as a pseudo-random normal random variable with mean as in the second column of Table 1 and standard deviation as in the last column. Event B was ignored because there is insufficient information given for us to derive a distribution for its date of occurrence. For each simulated sequence, the interevent times were computed and fifty simulations generated for the parameters of each of the exponential and lognormal recurrence time models, following the procedure of section 7.1 above. The hazard for each simulation of the event time data was computed by mixing the distributions as in equation 7 over the simulated parameter values and the hazard for each model computed by mixing the hazards over the simulated event time

data sets (as in equation 7). The results are shown in Figures 9–14.

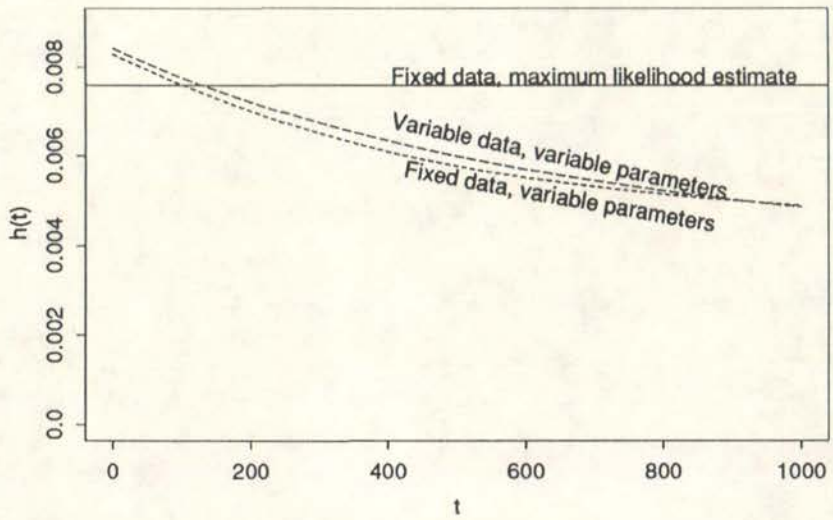

Figure 9: Pallett Creek hazard functions based on the exponential model for recurrence intervals. Time t is reckoned in years from the time of the last fault movement. The hazard $h(t)$ is measured in events per year. The recommended curve is that for variable data, variable parameters.

In Figures 9 and 12 our variable data, variable parameter estimates of hazard are compared to two previously published estimation methods: the common fixed data, maximimum likelihood method, which does not account for uncertainty in either the data or the parameters, and the fixed data, variable parameter method, similar to that proposed by Davis *et al.*, which allows for uncertainty in the parameters but not in the data. Davis *et al.* also produced an estimate which purported to allow for uncertainties in the data as well as the parameters but did this by assuming the interevent times were independent (clearly untrue) and made no apparent distinction between the hazard mixing which is appropriate for data uncertainties and the distribution mixing which is appropriate for parameters. Our variable data, variable parameter

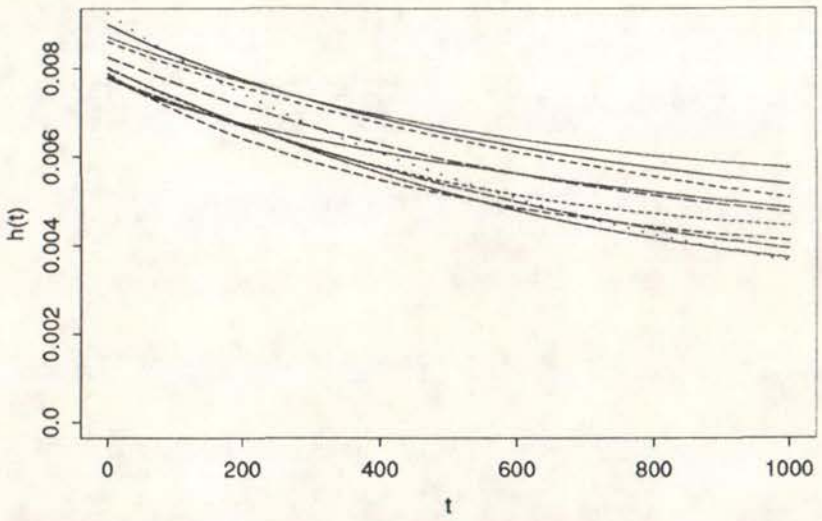

Figure 10: Pallett Creek hazard functions based on the exponential model for different randomly sampled data sets allowing for uncertainty in parameter values.

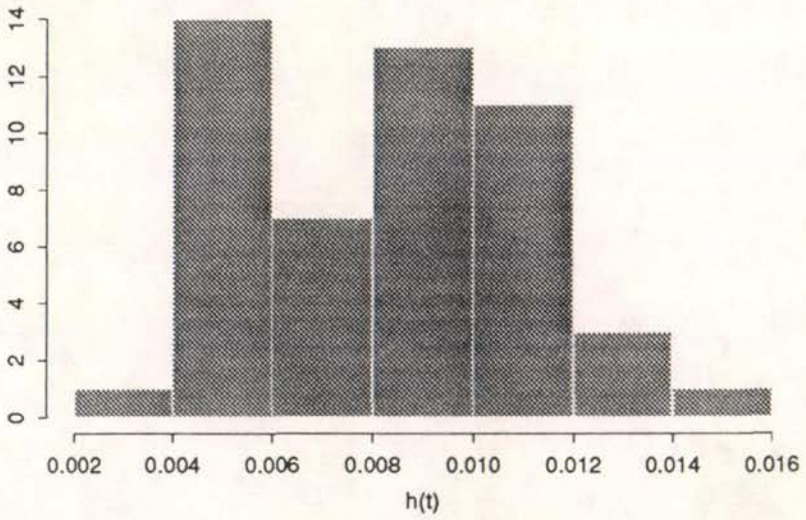

Figure 11: Histogram of exponential hazards for different randomly sampled parameter values, using central estimates of rupture times at Pallett Creek.

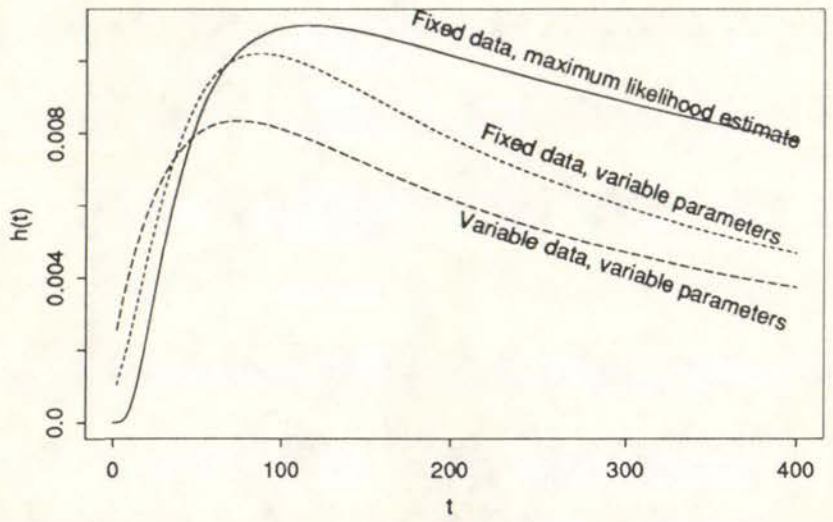

Figure 12: Pallett Creek hazard functions based on the lognormal model. The recommended curve is that for variable data, variable parameters.

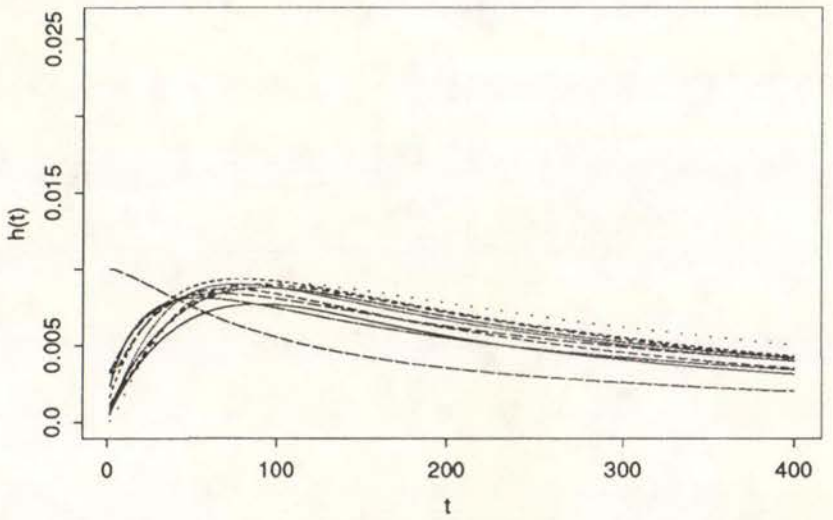

Figure 13: Pallett Creek hazard functions based on the lognormal model for different randomly sampled data sets allowing for uncertainty in parameter values.

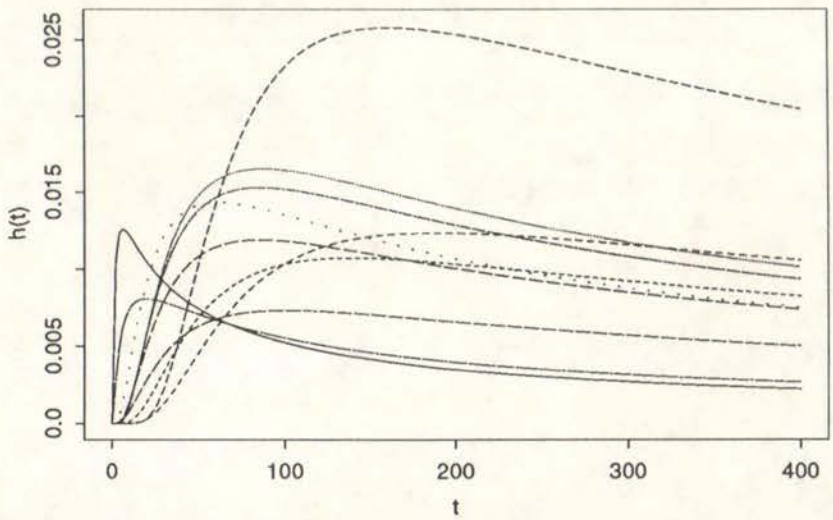

Figure 14: Pallett Creek hazard functions based on the lognormal model for different randomly sampled parameter values using a fixed data set.

estimate is thus more satisfactory in its handling of data uncertainties, but still can be improved upon by a more careful and detailed interpretation of the geological information and its uncertainties. The important thing to realise is that our general approach permits such detailed information on distributions and their interdependencies to be incorporated into the estimates of hazard; there is no longer the need to oversimplify the data distributions to conform to ideal statistical models.

In Figure 10 the sensitivity of the hazard (under the exponential model) to uncertainties in the data is displayed by plotting the hazard functions for different randomly sampled data sets. Figure 11 shows the sensitivity to parameter uncertainty. In this case, since the individual hazard functions are constant for fixed parameters, the variability is better displayed by a histogram. Figures 13 and 14 show the corresponding sensitivity analyses for the lognormal model. It is notable

that, in the Pallett Creek case, the sensitivity of the hazard to parameter uncertainties is greater than the sensitivity to data uncertainties. This can be seen by the relatively wide spread of hazard levels in Figures 11 and 14 compared to the spread in Figures 10 and 13 respectively. It is also interesting to compare the recommended curve for the exponential model in Figure 9 with that for the lognormal model in Figure 12. The lognormal curve starts at a lower level (actually zero) and rises to a higher value than the exponential curve between $t = 50$ and $t = 100$, but then drops away more rapidly. The dropping hazard under the exponential model (Figure 9) is due to adjustment to the parameter distribution as the elapsed time since the last earthquake increases.

Another feature is that the differences between the exponential and lognormal curves calculated according to the recommended method to take account of uncertainties in data and parameters are relatively small compared to the differences between the curves calculated with fixed data and maximimum likelihood estimates (Figures 9 and 12). This is reflected also in the rather similar probabilities for these two cases in Table 2, which shows conditional probabilities of rupture within the the next 50, 100, 200 and 300 years from 1990 A.D.

Table 2
Conditional probability of rupture from 1990 A.D. at Pallett Creek

	years	50	100	200	300
Exponential Model	Fixed data, maximum likelihood	.32	.53	.78	.90
	Fixed data, variable parameters	.30	.51	.75	.86
	Variable data, variable parameters	.31	.52	.76	.87
Log normal Model	Fixed data, maximum likelihood	.41	.64	.86	.94
	Fixed data, variable parameters	.36	.56	.77	.86
	Variable data, variable parameters	.30	.48	.68	.78

8.2 The Wellington fault

Data from the Wellington-Hutt Valley segment of the Wellington fault can be summarised as follows (Berryman, 1990; Van Dissen *et al.*, 1992):

- The age and displacement of the oldest dated terrace (near Upper Hutt) are 140 ± 10 ka and 940 ± 40m, respectively.
- Two younger terraces have ages and displacements of 14 ± 4 ka, 104 ± 10 m, and 70 ± 5 ka, 437 ± 20 m respectively.
- Along the same portion of the fault, single event displacements range from 3.2 to 4.7 m.

- The most recent rupture event is estimated to have occurred between 340 and 490 years ago.

Berryman (1990) states that the above uncertainties regarding the terrace ages and displacements approximate 95% confidence intervals. In what follows they are regarded as ±2 standard deviation limits of normally distributed estimates. Similarly the interval for the single event displacements is regarded as a ±2 standard deviation limit for the mean single event displacement, assumed normally distributed.

Only the most recent two events have been approximately dated on this segment, so it is not possible to estimate directly the coefficient of variation of interevent times. However the reasonably consistent size of single-event displacements measured at Te Marua, and the relatively constant slip rate calculated at Emerald Hill gives us reason to believe that the earthquake recurrence time on the Wellington–Hutt Valley segment of the Wellington fault is likely to be somewhat regular rather than purely random; a coefficient of variation less than 1 thus seems likely. In what follows the distribution for the coefficient of variation is "borrowed" from the Pallett Creek data[2], i.e.

$$f(\sigma) \propto \sigma^{-n} \prod_{i=1}^{n} \exp[-\frac{1}{2}(\frac{\log T_i - \hat{\mu}}{\sigma})^2] \tag{30}$$

where $T_i, i = 1, \ldots, n$ are the Central estimates of the interevent times at Pallett Creek and $\hat{\mu}$ is the maximum likelihood estimate of the lognormal parameter μ using these central estimates.

Again the hazard is estimated both for the exponential and lognormal recurrence time distributions. For illustration we demonstrate and compare the results of two different approaches using different data. The procedures are as follows.

Method A:

1. Generate 50 pseudo random samples from the distributions for age of the oldest dated terrace $\sum T_i$ (where T_i now refer to the

[2]The maximum likelihood estimate of the coefficient of variation at Pallett Creek was 0.84

interevent times on the Wellington fault), the corresponding total displacement D_T, and the mean single event displacement D_1. In the case of the lognormal model we also generate 50 random samples from the distribution for coefficient of variation described above.

2. For each sample pair (D_T, D_1), estimate the number of displacements n as the integer nearest to D_T/D_1. Given $\sum T_i$ and n (which is quite large, being of order 200), the uncertainty of the exponential parameter, and of the remaining lognormal parameter, is small and therefore ignored.

3. In the case of the exponential model estimate the rate parameter μ by $n/\sum T_i$ for each sample pair $(\sum T_i, n)$; this coincides with the (constant) value of the hazard function. Estimate the unconditional hazard by the average over the 50 samples.

4. In the case of the lognormal model, for each sampled triple $(\sum T_i, n, \sigma)$ estimate the parameter γ by $\sum T_i/n$ and hence calculate $\mu = \log(\gamma) - \frac{1}{2}\sigma^2$. Hence calculate the expected hazard function $E[h(t|a < t < b; \mu, \sigma)]$ using equation 25, where a refers to 340 years ago and b to 490 years ago. Calculate the unconditional hazard function as the mean over the 50 samples.

The results for method A are seen in Figures 15–17. Figure 15 is a histogram of the (constant) hazard functions obtained for 50 samples of $\sum T_i$ and n using the exponential model. Figure 16 shows the hazard functions obtained from 50 samples of $\sum T_i$, n and σ using the lognormal model and Figure 17 shows the result of averaging them. Overall the hazard is higher under the lognormal model for the period shown because the time since the last event is much less than the estimated mean recurrence interval. The hazard is fairly constant under the lognormal model for the period plotted because the full width of the time-scale is only a fraction of the mean recurrence interval.

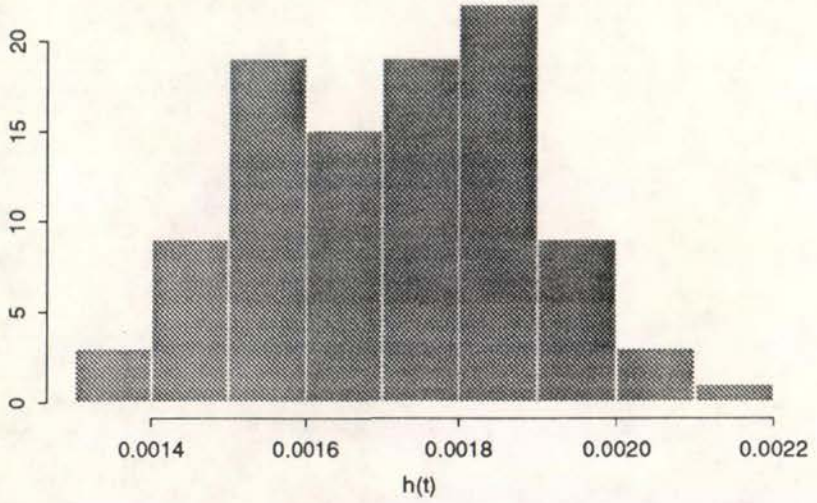

Figure 15: Histogram of Wellington fault hazards based on exponential model (Method A), for different randomly sampled data sets.

In the second approach we use the age and displacement data from the youngest of the three dated terraces. Since the value of n is much smaller than in Method A the parameter uncertainties are larger and cannot be ignored. Also it is not possible to use equation 25 in the lognormal case to handle the uncertainty in the timing of the most recent event because the mixture of distributions approach must be used for μ and the mixture of hazards approach for the time since the last event. In detail, the second approach is as follows.

Method B:

1. Generate 50 pseudo random samples from the distributions for age of the youngest dated terrace $\sum T_i$, the corresponding total displacement D_T, the mean single event displacement D_1 and the time of the last event τ (as uniform on the period from 340 to 490 years ago). In the case of the lognormal model generate also 50 random samples from the distribution for coefficient of variation as described above.

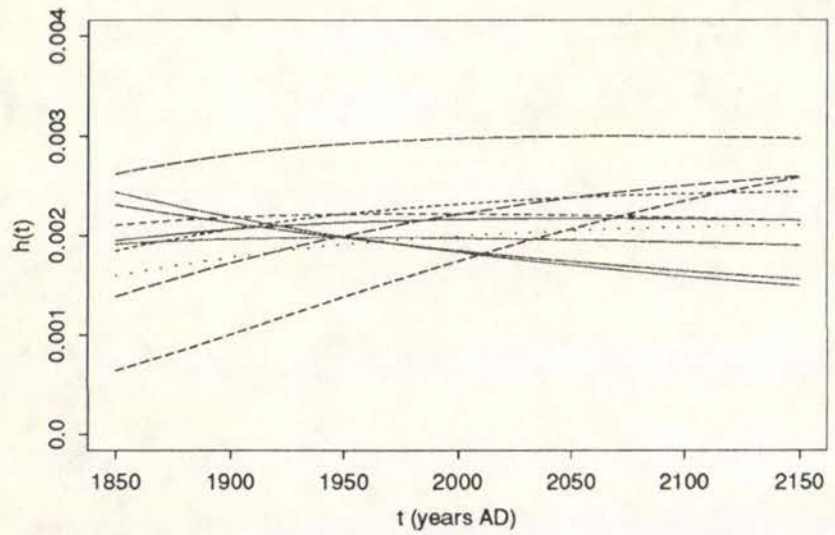

Figure 16: Wellington fault hazard variation with time based on the lognormal model (Method A), for different sampled triples $(\sum T_i, n, \sigma)$.

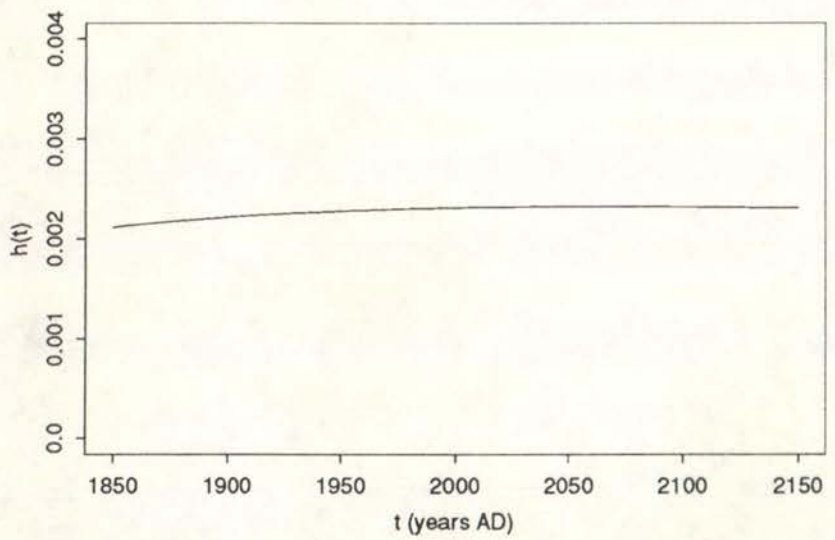

Figure 17: Wellington fault hazard variation with time using the lognormal model (Method A), averaged over sampled data sets.

2. For each sample pair (D_T, D_1), calculate the number of displacements n as the integer nearest to D_T/D_1.

3. In the case of the exponential model, for each sample triple $(\sum T_i, n, \tau)$ generate 50 samples for the the rate parameter μ and calculate the hazard $h(t|\sum T_i, n)$ by a mixture of distributions over the values of μ. Determine the unconditional hazard function as the average over the 50 samples.

4. In the case of the lognormal model, for each sampled quadruple $(\sum T_i, n, \sigma, \tau)$ generate 50 samples of the lognormal parameter μ using the gamma distribution approximate method described above. Calculate the hazard function $h(t|\sum T_i, n, \sigma, \tau)$ using a mixture of distributions over values of μ. Determine the unconditional hazard function as the mean over the 50 samples of. $\sum T_i, n, \sigma$ and τ.

The results for method B are shown in Figures 18–23. In this case the effect of taking account of uncertainties in both data and parameters is seen to be quite small, at least over the time period for which the hazard functions are plotted. The hazard under method B is generally slightly higher than under method A, due to the slightly faster estimated slip rate associated with the youger dated terrace. Again the hazard is somewhat higher for the lognormal model than for the exponential model but the difference is not great when the imponderable uncertainties associated with the assumptions necessary for recurrence-time modelling (discussed in section 4) are considered.

The conditional probabilities of rupture within the next 50, 100, 200 and 300 years from 1990 A.D. under each method are tabulated in Table 3.

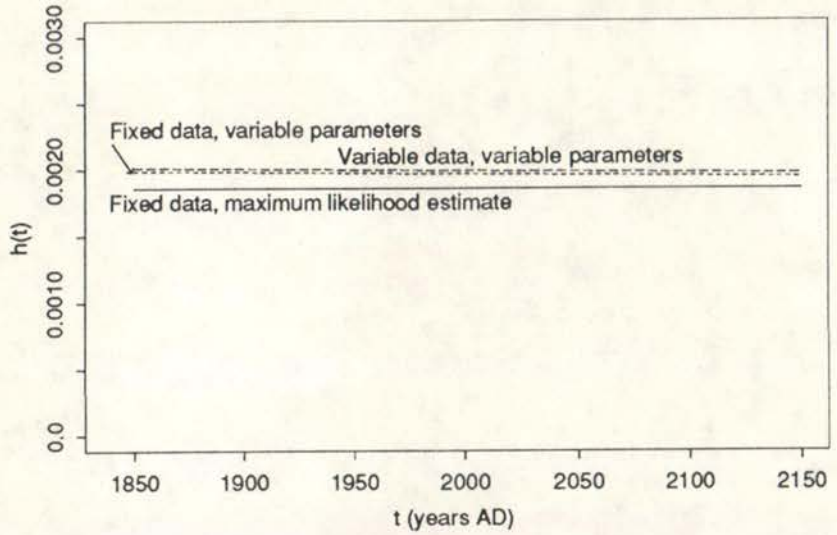

Figure 18: Wellington fault hazard variation with time based on the exponential model (Method B).

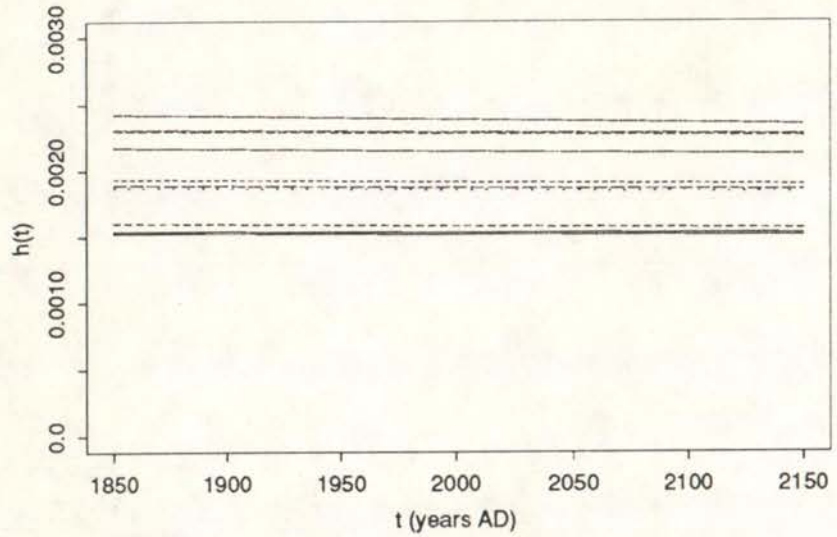

Figure 19: Wellington fault hazard variation with time based on the exponential model for different randomly sampled data sets allowing for uncertainty in parameter values (Method B).

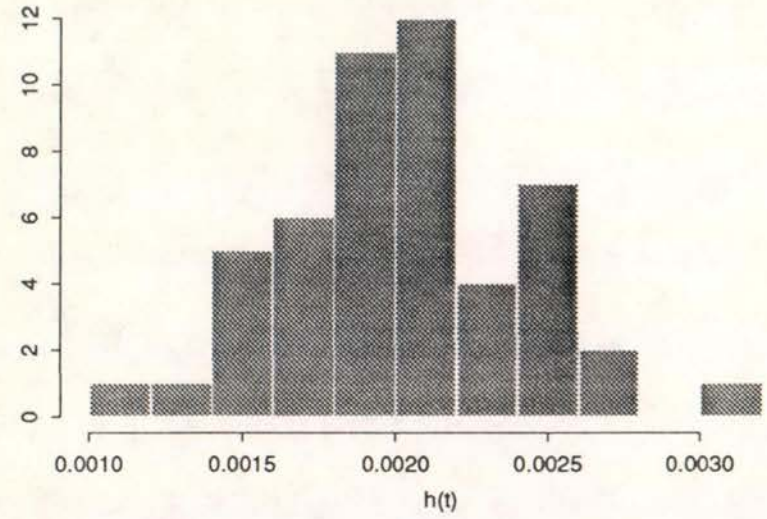

Figure 20: Histogram of exponential hazards for different randomly sampled parameter values, using fixed data (Method B).

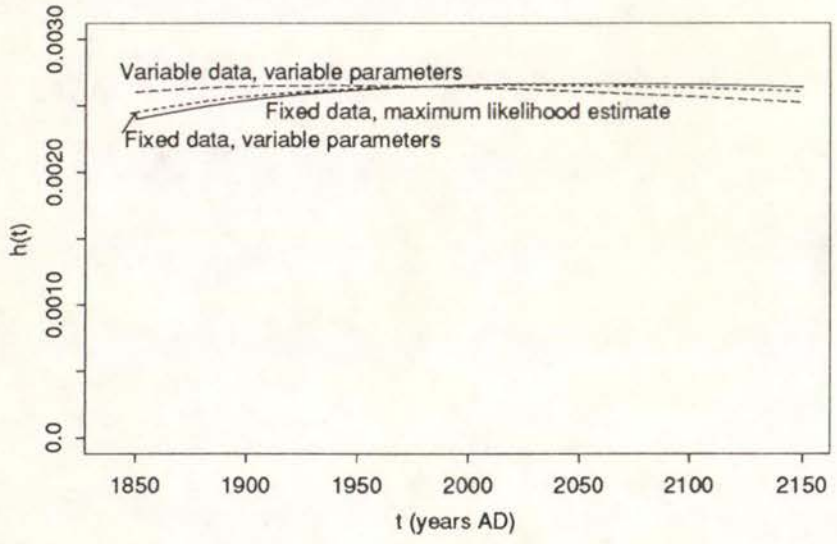

Figure 21: Wellington fault hazard variation with time based on the lognormal model (Method B).

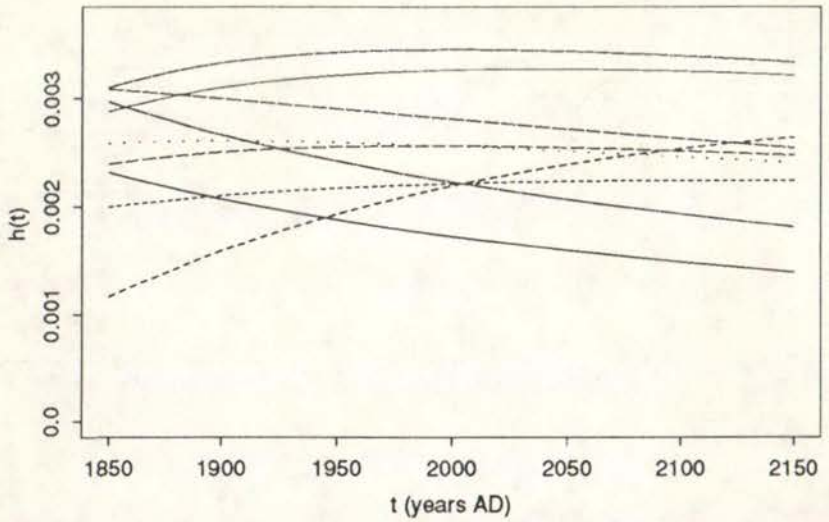

Figure 22: Wellington fault hazard variation with time based on the lognormal model for different randomly sampled data sets allowing for uncertainty in the parameter values (Method B).

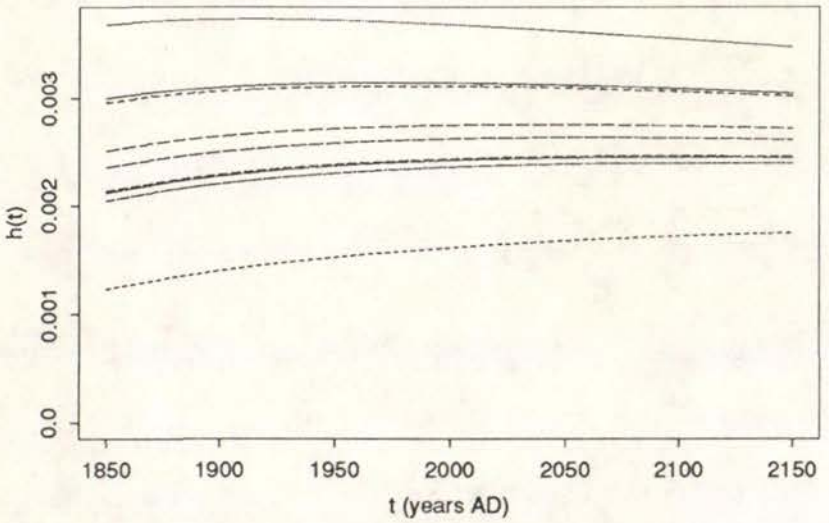

Figure 23: Wellington fault hazard variation with time based on the lognormal model for different randomly sampled parameter values using a fixed data set (Method B).

Table 3

Conditional probability of rupture from 1990 AD for Wellington.

	years	50	100	200	300
Method A Exponential Model	Variable data Variable c.v.	.08	.16	.29	.40
Log normal Model	Variable data Variable c.v.	.11	.21	.37	.50
Method B Exponential Model	Fixed data, maximum likelihood	.09	.17	.31	.43
	Fixed data, variable parameters	.09	.18	.32	.44
	Variable data, variable parameters	.10	.18	.33	.45
Log normal Model	Fixed data, maximum likelihood	.12	.23	.41	.55
	Fixed data, variable parameters	.12	.23	.41	.54
	Variable data, variable parameters	.12	.23	.40	.53

Figure 24 shows the preferred hazard curves, allowing for variable data and variable parameters, under each method and model for comparison.

Neither of the above estimations for the Wellington fault using the lognormal model can be regarded as a definitive estimation of hazard for the Wellington fault; the question of what is the real distribution for the coefficient of variation cannot be answered from existing data. The use of the lognormal model in such circumstances is scientifically

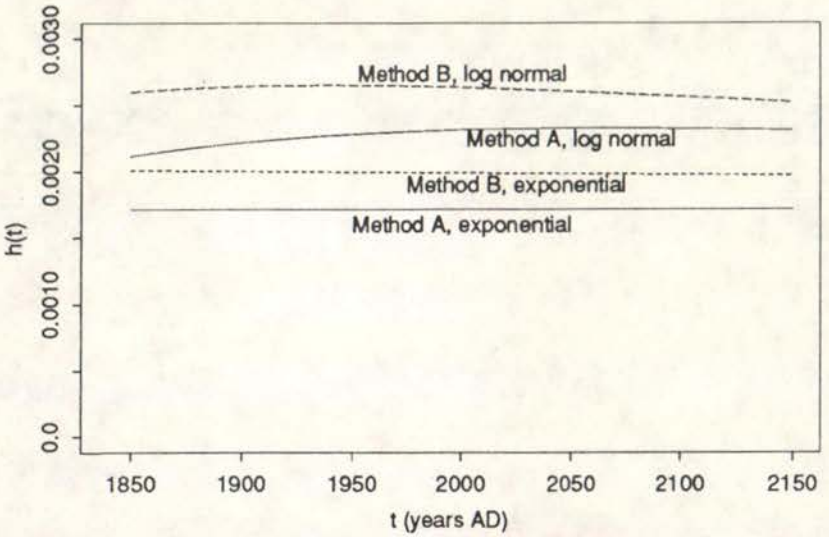

Figure 24: Wellington fault hazard functions allowing for variable data and variable parameters under each of Method A and Method B and based on exponential and log normal models

relevant only to the question of sensitivity. It is legitimate to ask: what if the coefficient of variation is the same as for some other fault, or the same as in the generic distribution of Nishenko and Buland (1987), to see if the hazard is materially different from that under the exponential model. If there is no practical difference between the two models for the period of interest, then the coefficient of variation is a matter of no practical interest for that particular fault. If it turns out that there is a difference, then the extent to which estimation of hazard based on an assumed coefficient of variation should be given credence is a matter of expert judgement and debate.

9 Implications for future geological studies

This study has implications for the way in which data should be presented in geological studies related to fault-zone hazard. It is important that uncertainties in estimated quantities such as ages, slip-rates and displacements are reported fully so that the best use can be made of this type of data in future hazard studies for specific sites, or regions, or for earthquake loadings code purposes. In some cases, where estimates of fault rupture times and displacements are given, the information on data uncertainties is insufficiently detailed to allow these methods to be properly applied. This is particularly so when the timing of past ruptures has not been directly determined. Considering the effort and cost that goes into an active fault investigation, skimping on the description of uncertainties cannot be justified. Field geologists should seek expert statistical advice when nonstandard situations arise in the estimation of any of the above quantities.

10 Conclusion

The study enables important improvements to the quality of hazard assessment (in fault zones) based on geological information. The use of all relevant information, including that on uncertainties, in applying particular models will lead to more soundly-based comparisons between models and more informed assessments of model performance. Bringing together all the uncertainties into a single estimate, rather than a set of conditional estimates, allows easier interpretation of the results and, again, better comparisons between alternative models.

The examples given here have been purely for the purposes of illustration of the general method, and should not be regarded as definitive estimates of hazard.

Acknowledgements

This study was substantially funded by a research grant from the New Zealand Earthquake and War Damage Commission. Miss S.M. Harper carried out the computations and graphical analyses.

References

Berryman, K.R. (1990). Late Quaternary movement on the Wellington Fault in the Upper Hutt area, New Zealand. *N.Z. J. Geol. Geophys. 33*:257–270.

Brillinger, D.R. (1982). Seismic Risk Assessment: some statistical aspects. *Earthqu. Predict. Res. 1:* 183–185.

Davis, P.M., Jackson, D.D., and Kagan, Y.Y. (1989). The longer it has been since the last earthquake, the longer the expected time till the next? *Bull. Seism. Soc. Am. 79:* 1439–1456.

DeGroot, M.H. (1963). *Optimal Statistical Decisions*. McGraw-Hill, New York, 489pp.

Jacob, K.H. (1984). Estimates of long-term probabilities for future great earthquakes in the Aleutians. *Geophys. Res. Lett., 11:* 295–298.

Johnson, N.L. and Kotz, S. (1970). *Distributions in Statistics: Continuous Univariate Distributions - 1.* Houghton Mifflin, Boston, 300pp.

Kiremidjian, A.S. and Anagnos, T. (1984). Stochastic slip-predictable model for earthquake occurrence. *Bull. Seismol. Soc. Am., 74*: 739–755.

Nishenko, S.P., (1985). Seismic potential for large and great interplate earthquakes along the the Chilean and South

Peruvian margins of South America - a quantitative reappraisal. *J. Geophys. Res., 90:* 3589–3616.

Nishenko, S.P. (1991). Circum-Pacific seismic potential: 1989–1999. *PAGEOPH, 135:* 169–259.

Nishenko, S.P and Buland, R. (1987). A generic recurrence interval distribution for earthquake forecasting. *Bull. Seism. Soc. Am., 77:* 1382–1399.

Pearson, G.W. and Stuiver, M. (1986). High precision calibration of the radiocarbon time scale. *Radiocarbon, 28:* 839–862.

Rhoades, D.A. (1989) Theory of multiple precursors: variation of hazard in time and space. *Proceedings of the 4th International Symposium on the Analysis of Seismicity and Seismic Risk, Bechyne Castle, September 4–9, 1989*: 454–461, Czechoslovak Academy of Sciences.

Rhoades, D.A. and Evison, F.F. (1989). Time-variable factors in earthquake hazard. *Tectonophysics, 167:* 201–210.

Rhoades, D.A. and Millar, R.B. (1983). Estimating the hazard of surface faulting in a single fault zone. *Seismotectonic Hazard Evaluation of the Clyde Dam Site*, NZ Geological Survey, 1983, Part II, Appendix 10.

Sieh, K., Stuiver, M. and Brillinger, D. (1989). A more precise chronology of earthquakes produced by the the San Andreas fault in Southern California. *J. Geophys. Res., 94:* 603–623.

Stuiver, M. and Pearson, G.W. (1986). High precision calibration of the radiocarbon time scale. *Radiocarbon, 28:* 805–838.

Van Dissen, R.J., Berryman, K.R., Pettinga, J.R., and Hill, N.L. (1992). Paleoseismicity of the Wellington–Hutt Valley

Appendix

Proof of Equation 25

$$E[h(t)|a < t < b] = \int_a^b h(t)p(t)dt$$

where $p(t)$ is the probability density for the time elapsed since the last movement given the cumulative distribution function $F(t)$ for the recurrence time. Since the probability that the drought exists for time t is $1 - F(t)$, $p(t) \propto 1 - F(t)$. Hence

$$\begin{aligned} E[h(t)|a < t < b] &= \frac{1}{\int_a^b [1 - F(t)]dt} \int_a^b h(t)[1 - F(t)]dt \\ &= \frac{F(b) - F(a)}{\int_a^b [1 - F(t)]dt}. \end{aligned}$$

If F is lognormal(μ, σ), then, making a change of variable to $u = (\log t - \mu)/\sigma$ and integrating by parts we have, for the denominator,

$$\begin{aligned} \int_a^b [1 - F(t)]dt &= \int_{\frac{\log a - \mu}{\sigma}}^{\frac{\log b - \mu}{\sigma}} [1 - \Phi(u)]\sigma e^{\mu + \sigma u} du \\ &= b[1 - \Phi(\frac{logb - \mu}{\sigma})] - a[1 - \Phi(\frac{loga - \mu}{\sigma})] + \int_{\frac{\log a - \mu}{\sigma}}^{\frac{\log b - \mu}{\sigma}} \phi(u)e^{\mu + \sigma u} du \end{aligned}$$

where ϕ is the standard normal density. The last integral on the right hand side can be written as

$$\begin{aligned} & \frac{e^{\mu}}{\sqrt{2\pi}} \int_{\frac{\log a - \mu}{\sigma}}^{\frac{\log b - \mu}{\sigma}} e^{-\frac{1}{2}u^2} e^{\sigma\mu} du \\ = & \frac{e^{\mu + \frac{1}{2}\sigma^2}}{\sqrt{2\pi}} \int_{\frac{\log a - \mu}{\sigma}}^{\frac{\log b - \mu}{\sigma}} e^{-\frac{1}{2}(u - \sigma)^2} du \\ = & e^{\mu + \frac{1}{2}\sigma^2} [\Phi(\frac{\log b - \mu - \sigma^2}{\sigma} - \frac{\log a - \mu - \sigma^2}{\sigma})] \end{aligned}$$

where the last step is the result of a change of variable to $v = u - \sigma$. Collecting up terms in the denominator, and noting that the numerator

is given by

$$F(b) - F(a) = \Phi(\frac{\log b - \mu}{\sigma}) - \Phi(\frac{\log a - \mu}{\sigma}),$$

Equation 25 follows.