Validation of using Gumbel probability plotting to estimate Gutenberg-Richter seismicity parameters

Mike Turnbull[1] and Dion Weatherley[2]

[1] Central Queensland University
[2] Queensland University

Abstract:

The Gumbel Type I statistics of extreme events have been successfully used in the past to forecast various natural events such as annual exceedence of design flood level, and hail fall. Some attempts have been made to determine seismicity parameters using the annual maximum magnitude events in historic records. The results from these determinations have invariably been criticized for various reasons, including the perception that the methodology ignores important data, and that the method has no verification basis. This paper addresses both topics by discussing the principles of the Gumbel Type I statistical method, and verifying that the method is capable of reliably deducing the Gutenberg-Richter seismicity parameters of complete synthetic earthquake calendars, using only the annual maxima.

Introduction

It is common to characterize temporal and quantitative earthquake seismicity of a region by respectively specifying values for the a and b parameters of the Gutenberg-Richter seismicity model (the G-R model). Estimations of these parameters can be derived from a number of statistical processes. In situations where a comprehensively complete catalogue of earthquake events is not available, methods provided by the statistics of extreme events (the so-called extreme value theory (EVT)) have been applied, using reduced variate probability plotting.

The generalized EVT cumulative distribution function (cdf) reduces to one of three specific Fisher Tippett distributions (Fisher & Tippett, 1928), depending on the value chosen for its three parameters, ξ, $\theta(>0)$, and $k(>0)$. These three distributions are summarized below (Johnson et al, 1995).

Fisher Tippett Type 1:

$$\Pr[X \le x] = \exp\{-\exp\{-1/\theta(x-\xi)\}\} \quad \ldots \quad \text{Eq. 1}$$

Fisher Tippett Type 2:

$$\Pr[X \le x] = 0, \quad \text{where } x < \xi \quad \ldots \quad \text{Eq. 2}$$
$$= \exp\{-\exp\{-(1/\theta(x-\xi))k\}\}, \quad \text{where } x \ge \xi$$

Fisher Tippett Type 3:

$$\Pr[X \le x] = \{-\exp\{-(1/\theta(\xi - x))k\}\}, \quad \text{where } x \le \xi \quad \ldots \quad \text{Eq. 3}$$
$$= 1, \quad \text{where } x > \xi$$

The Type 2 distribution is often referred to as the Fréchet distribution. The Type 3 distribution is often referred to as the Weibull distribution. The Type 1 distribution is mostly referred to as the Gumbel distribution, but is sometimes referred to as the log-Weibull distribution. In this paper it will be referred to as the Gumbel distribution.

There are two common criticisms made, arguing that the probability plotting method of analysing extreme events to estimate regional seismicity is of little value to practical seismology. These criticisms are that:

1. The extreme value methods only assess the few maximum value events and ignore the many other important smaller events.
2. The various methods of determining the plotting positions used to calculate the reduced variate are arbitrary in nature. Therefore the choice of plotting position algorithm can be used to manipulate the results.

This paper addresses these two criticisms via counter-arguments and a demonstration that the reduced variate probability plotting method in conjunction with Gumbel statistics of extreme events can reproduce accurate estimates of a priori seismicity parameters used to generate synthetic earthquake calendars. Our analysis consists of two parts. Firstly we demonstrate that the probability plotting method estimates to within 2% accuracy, the Gumbel parameters of a synthetic dataset constructed with a priori values of these parameters. Secondly we apply the Gumbel method to analyse synthetic seismicity calendars generated from a Gutenberg-Richter distribution with prescribed a- and b-values. Our results testify that the Gumbel method accurately estimates the a and b values of the underlying G-R source distribution, via statistical analysis of only the extreme values of the synthetic catalogues.

Are important data being ignored?

Statistical analysis aims to provide an accurate model for a given set of observations, using some assumptions about the underlying process giving rise to the observations. In the case of regional seismicity, one assumes the underlying process gives rise to a Gutenberg-Richter frequency-magnitude distribution: a two-parameter model determining the average rate of seismicity (a value) and the scaling of recurrence intervals with given earthquake magnitude (b value). For a particular region, one aims to estimate the values for these two parameters via curve fitting of the observed historical seismicity. Since the dataset of observations is invariably only a small subset (or sampling) of the seismic history and the observations may contain errors (e.g. imprecise magnitude determination or poor detection level) one cannot expect to obtain an arbitrarily accurate estimation of the model parameters.

It is well-known that estimated values for the model parameters may be significantly skewed when using a dataset which does not provide a sample a data set containing adequate samples of the full range of observable values. Seismicity particularly suffers from this limitation as historical seismic catalogues are typically complete for large magnitudes (the extreme values of the G-R distribution) but incomplete or non-existent for smaller magnitudes. Historical catalogues are biased towards extreme values.

The Fisher-Tippett probability distributions are specifically formulated to model the extreme data values that are invariably found in samples extracted from underlying source distributions. EV distributions provide a parameterisation for the extreme values that is related to the parameters of the source distribution, while taking into account the inherent bias towards extreme values in the dataset under analysis. It was Fisher and Tippet (1928) who proved that no matter what source probability distribution data is derived from, the distribution of extreme data values will necessarily converge to one of the three forms Eq. 1, 2 or 3.

The perception that extreme value methods ignore important small value data is false. EV methods are designed to model the distribution of extreme values accurately, not the distribution of non-extreme values. Including these latter values in the analysis would be erroneous. Since the dataset of extreme values is complete, one does not suffer from the finite sampling issues when estimating the parameters of the EV distribution. It must be emphasised that EV methods make allowance for the bias towards extreme values in the original dataset. This is codified in the relationship between EV model parameters and those of the source distribution. Thus it is possible, by analysing a catalogue of extreme values, to accurately estimate the parameters of the source distribution. Given the indisputable bias towards large magnitudes in seismic catalogues, EV methods are well-suited for modelling regional seismicity.

The probability integral transformation theorem

The theorem of probability integral transformation states that any cumulative distribution function, considered as a function of its random variable X, is itself a uniform random variable on the closed interval (0,1) (Bury, 1999, p 25).

$F(X; \theta) = U$ … (Eq. 4)

where θ represents parameters, either known or not yet determined.

A consequence of this theorem is that all possible values of X are equally likely. So that any sample variate $F(x_i; \theta)$ derived from the parent distribution $F(X; \theta)$ can be expressed in the form:

$F(x_i; \theta) = u_i$ … (Eq. 5)

where u_i is a value in the closed interval (0, 1), and where all values of u_i are equally likely.

A corollary of the probability integral transformation theorem is that:

$x_i = F\text{-}1(u_i; \theta)$ (Eq. 6).

This corollary has two important applications in practice – simulated random observations, and probability plotting.

Simulating random variates

The corollary of the probability integral transformation theorem provides a means of simulating random variates from any known probability distribution. By substituting random numbers u_i from the closed (0, 1) interval into the inverse of the distribution's cumulative distribution function, independent identically distributed random variates can be generated.

For example (Bury, 1999, p 268), the cdf of the Gumbel distribution may be expressed in the form

$F(x; \mu, \sigma) = \exp\{-\exp\{-1/\sigma\ (x - \mu)\}\} = u$ (say). (Eq. 7)

By inversion

$$x = \mu - \sigma\ln(-\ln(u)) \quad \text{(Eq.8)}$$

Therefore, simulated random variates x_i from the Gumbel distribution can be generated using the following formula, where u_i is a random number on the closed interval (0, 1).

$$x_i = \mu - \sigma\ln(-\ln(u_i)) \quad \text{(Eq.9)}$$

Probability plotting

Manipulation of Eq. 9 produces the following linear relation.

$$-\ln(-\ln(p_i)) = 1/\sigma\ (x_i - \mu) \quad \text{(Eq.10)}$$

where the u notation has been replaced by a p, for reasons that will become clear below.

This relation provides a potential means of testing whether a set of n experimental observations $\{x_i\}n$ is a sample from a Gumbel distribution. If the reduced variates $\{-\ln(-\ln(p_i))\}n$ are plotted against the experimental observations $\{x_i\}n$, and a straight line graph results, then the postulated Gumbel parent distribution is confirmed, and ordinary linear regression can be used to estimate the parameters σ and μ from the slope and intercept. There is one difficulty in accomplishing this task. In any real experimental situation the observations $\{x_i\}n$ are known, but the n reduced variates $\{-\ln(-\ln(p_i))\}n$ cannot be calculated exactly because the plotting positions $\{p_i\}n$ are unknown.

The only things that can be assumed regarding the p_i values is that they are in the closed interval (0, 1), and that each value has equal likelihood of presence. This information suggests a widely used, but controversial, method for producing artificial plotting positions that can be substituted for the actual ones. The method used to determine the substitute plotting positions can be described as follows.

The n observations are first ordered and ranked according to their relative values. Depending on the requirements of the particular situation this ranking may be in

ascending or descending order. The examples described here will use ascending order. The ordered observations are notated as

$\{x_i\}^*n \_ \{x_1 \le x_2 \le x_3 \le \ldots \le x_{n-2} \le x_{n-1} \le x_n\}$

where x_1 is the smallest valued variate, x_n is the largest valued variate, and the subscript values are the variate ranks.

The next step is the controversial part of the method. The rank value of the mth ordered variate is used to determine an artificial plotting position quantile p_m for that variate. There is no single definitive formula or equation for doing this. However, there are guidelines for doing so.

Gumbel (1958) expressed the following five conditions as requirements that substitute plotting positions should necessarily fulfil.

1. The plotting position should be such that all observations can be plotted.
2. The plotting position should lie between the observed frequencies (m – 1)/n and m/n and should be universally applicable, i.e., it should be distribution-free. This excludes the probabilities of the mean, median, and modal mth value which differ for different distributions.
3. The return period of a value equal to or larger than the largest observation, and the return period of a value smaller than the smallest observation, should approach n, the number of observations. This condition need not be fulfilled by the choice of the mean and median mth value.
4. The observations should be equally spaced on the frequency scale, i.e., the difference between the plotting positions of the (m + 1)th and the mth observation should be a function of n only, and independent of m. This condition … need not be fulfilled for the probabilities at the mean, median, or modal mth values.
5. The plotting position should have an intuitive meaning, and ought to be analytically simple. The probabilities at the mean, modal, or median mth value have an intuitive meaning. However, the numerical work involved is prohibitive [at the time of writing. Current computing capabilities now make these calculations routine].

The simplest approach is to assume that the value of the plotting position quantile is equal to its fractional position in the ranked list, m/n. This would assign the quantile 1/n to the smallest plotting position and n/n = 1 to the largest. This is unsatisfactory because it leaves no room at the upper end for values greater than the largest variate observed thus far.

Most plotting position formulae are ratios of the form (m ± a)/(n ± b) where the addends and subtrahends are chosen to improve estimates in the extreme tails of the postulated distribution.

Gumbel (ibid) recommended the following quantile formulation, which calculates the mean frequency of the mth variate.

$$p_m = m / (n + 1) \quad \ldots \quad \text{(Eq. 11)}$$

This formulation ensures that any plotting position is as near to the subsequent one as it is to the previous. It also produces a symmetrical sample cdf in the sense that the same plotting positions will result from the data regardless of whether they are assembled in ascending or descending order.

A more sophisticated formulation is

$$p_m = (m - 0.3) / (n + 0.4) \quad \ldots \quad \text{(Eq. 12)}$$

This formulation approximates the median of the distribution free estimate of the sample variate to about 0.1% and, even for small values of n, produces parameter estimations comparable to the results obtained by maximum likelihood estimations (Bury, 1999, p 43).

Using the Gumbel distribution to model extreme earthquakes

Cinna Lomnitz (1974) showed that if an homogeneous earthquake process with cumulative magnitude distribution

$$F(m; \beta) = 1 - e^{-\beta m}; \quad m \geq 0 \qquad \text{(Eq. 13)}$$

is assumed (compare with Eq. 24), where β is the inverse of the average magnitude of earthquakes in the region under consideration; and α is the average number of earthquakes per year above magnitude 0.0; then y, the maximum annual earthquake magnitude, will be distributed according to the following Gumbel cdf.

$$G(y; \alpha, \beta) = \exp(-\alpha \exp(-\beta y)); \quad y \geq 0 \qquad \ldots \qquad \text{(Eq. 14)}$$

Using the probability integral transformation theorem, simulated maximum yearly earthquakes can be generated using the following inversion formula.

$$y_i = -(1/\beta) \ln((1/\alpha) \ln(1/u_i)) \qquad \ldots \qquad \text{(Eq. 15)}$$

The conversion factors to transform Eq. 4 and 6 to Eq. 11 and 12 are as follows.

$$\alpha = \exp(\mu / \sigma) \qquad \ldots \qquad \text{(Eq. 16)}$$
$$\beta = 1 / \sigma \qquad \ldots \qquad \text{(Eq. 17)}$$

Conversely:

$$\sigma = 1 / \beta \qquad \ldots \qquad \text{(Eq. 18)}$$
$$\mu = (1/\beta) \ln(\alpha) \qquad \ldots \qquad \text{(Eq. 19)}$$

Manipulation of Eq. 15 produces the following linear relation.

$$-\ln(-\ln(p_i)) = \beta y_i - \ln(\alpha) \qquad \ldots \qquad \text{(Eq. 20)}$$

where p represents the plotting position, and the left hand expression is the reduced variate that can be used to plot data that is postulated as being drawn from a Gumbel distribution.

Demonstration of Gumbel probability plotting

Eq. 15 was used to generate ten random, one thousand year catalogues of synthetic annual extreme earthquake magnitudes, using the input parameters α = 48, and β = 1.37. Each set of data was analysed by plotting Eq. 20, with the plotting positions determined using both Eq. 11 and 12. Figures 1a and 1b show one of the ten resulting graphical plots obtained using each plotting method. The dotted lines in Figure 1 are the ordinary linear regression approximations. The linear approximation equations and coefficients of linear determination (r^2) are shown at the top right hand corner of each graph. The visual interpretation of the graph is that, for the majority of the lower magnitude data, the Gumbel distribution is appropriate; but, for magnitudes above about

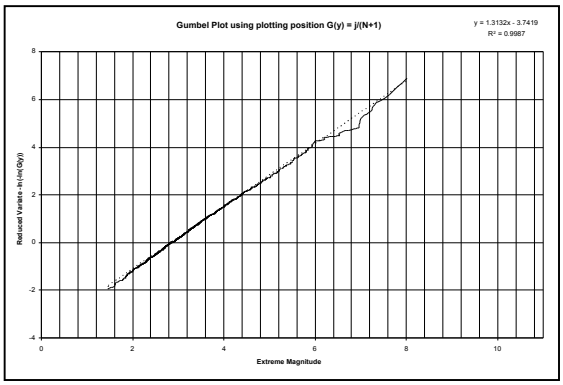

Figure 1(a): Gumbel Probability Plot using i/(n+1)

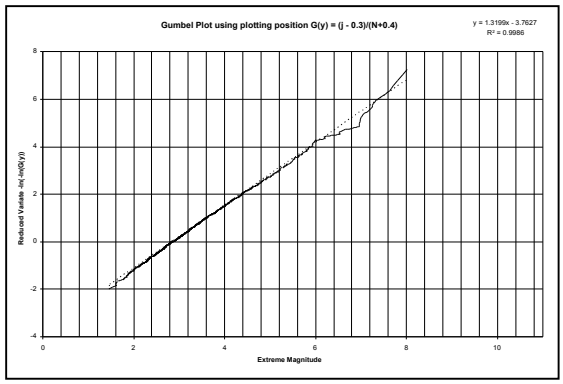

Figure 1(b): Gumbel Probability Plot using (i-0.3)/(n+0.4)

6.0 (i.e. for the extreme of the extreme values), the assumption of a Gumbel distribution may be suspect (in fact, a Weibull analysis may be more appropriate for those data).

Estimations of α and β were made using ordinary linear regression of each set of data. Table 1 summarises the resulting approximations, as well as showing the average and standard deviations of the estimated parameters.

Table 1: Parameter estimations using Gumbel Probability Plotting

Parameter estimation	_		_	
	pi=i/(n+1)	pi=(i-0.3) /(n+0.4)	pi=i/(n+1)	pi=(i-0.3) /(n+0.4)
Data Set 1	42.17857	43.06256	1.313218	1.319945
Data Set 2	44.00797	45.03307	1.328423	1.335858
Data Set 3	44.30699	45.27417	1.338465	1.345506
Data Set 4	51.20298	52.48388	1.409218	1.417366
Data Set 5	41.95601	42.79609	1.317657	1.324136
Data Set 6	49.80841	50.87322	1.374311	1.381216
Data Set 7	47.47786	48.46969	1.362259	1.369031
Data Set 8	50.64748	51.86245	1.382438	1.390142
Data Set 9	42.62512	43.62954	1.33023	1.337812
Data Set 10	43.24977	44.15485	1.348617	1.355476
Average	45.75	46.76	1.35	1.36
Std Dev	3.49	3.60	0.03	0.03
Std Error	1.10	1.14	0.0095	0.0095
Rel Error	2.4%	2.4%	0.7%	0.7%
Exact Value	48.00	48.00	1.37	1.37

It is evident that both plotting methods can estimate α and β within standard relative errors of 2.4% and 0.7% respectively, if sufficient trials are made. It is expected that trials with a larger number of data sets would improve the relative errors.

In real situations it may only be possible to extract a single useful data set from the earthquake history. This will limit the precision of parameter estimation in practice. For single estimations, there is a 95% confidence that α and β can be estimated within two standard deviations of the averages quoted in Table 1. That is, within 15% and 5% respectively.

It is clear from this demonstration that the fundamental method of probability plotting is scientifically sound in that it can reproduce accurate approximations of underlying process model parameters (at least for the two plotting position formulations used in this demonstration).

It is pointed out that the forgoing error analysis pertains to the method itself. Other errors in the inferred results of particular analyses may be introduced by faulty data. In particular incorrect determination of earthquake magnitudes may adversely affect inferred results.

Demonstration of Gutenberg-Richter parameter estimation using the Gumbel distribution

Background Theory

The Gutenberg-Richter (G-R) seismicity relation of earthquake frequency versus magnitude may be expressed as:

$$N(m \geq M) = 10^{(a - b\,m)} \quad \ldots \quad \text{(Eq. 21)}$$

where $N(m \geq M)$ is the number of earthquakes observed having magnitudes greater than or equal to M; and a and b are parameters specific to the observed data set. As a pragmatic mathematical and practical choice, the lower limit of M, M_0 is usually assigned the value zero. In that formulation the parameter a represents the logarithm to the base 10 of the number of independent earthquakes in the observation period with magnitude greater than or equal to zero.

$$a = \log_{10} N(m \geq M_0) \Rightarrow N(m \geq M_0) = 10^a \quad \ldots \quad \text{(Eq. 22)}$$

If it is assumed that all earthquake included in the data set are independent, and that each event has equal probability of occurring, then Eq. 21 can be normalised to produce a frequency relation as follows,

$$Pr(m \geq M) = N(m \geq M)/ N(m \geq M_0) = 10^{(a - b\,m)}\, 10^{-a} = 10^{-b\,m} \quad \ldots(\text{Eq. 23})$$

It can be seen from Eq. 23 that the value of the parameter b determines the propensity for lower or higher magnitude earthquakes. Smaller values of b model a system that has a greater propensity for larger magnitude earthquakes. It also demonstrates that magnitude of the earthquakes is not dependent on the a parameter. The cdf formulation is as follows.

$$Pr(m \leq M) = 1 - 10^{-b\,m} \quad \ldots \quad (\text{Eq. 24})$$

Using the probability integral transformation theorem, Eq. 24 can be inverted to produce a random magnitude generator

$$m = -1/b \log_{10}(1 - u) \quad \ldots \quad (\text{Eq. 25})$$

where u is a random number in the closed (0, 1) interval. Eq 25 also provides the reduced variate for conducting a G-R plot to test whether a data set is drawn from a G-R distribution.

If it is further assumed that the timing of the earthquake events is a Poisson process, then a random event generator can be devised (c.f. Bury, 1999, p 104).

$$t = -10^{-a} \ln(v) \quad \ldots \quad (\text{Eq. 26})$$

where t is a random time interval between events, v is a random number in the closed (0, 1) interval, and 10^{-a} is the average time between events.

From Eqs. 13 and 24, and the fact that α and 10^a specify the average time between events in the Gumbel and G-R formulations respectively, the relationships between the Gumbel parameters α and β and the G-R parameter a and b are seen to be

$$e^{-\beta} = 10^{-b} \quad => \quad b = \beta \log_{10} e \quad \ldots \quad (\text{Eq. 27})$$

$$\alpha = 10^a \quad => \quad a = \log_{10} \alpha \quad \ldots \quad (\text{Eq. 28})$$

Using the same parameter values that were employed in the demonstration of Gumbel plotting, if $\alpha = 48$, then $a \approx 1.69$; and if $\beta = 1.37$, then $b \approx 0.59$.

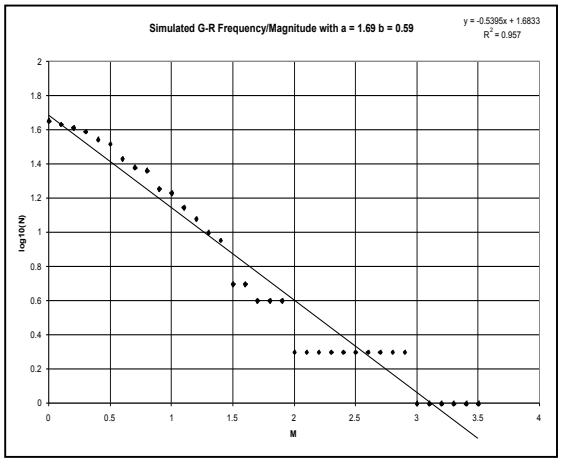

Figure 2(a) : G-R Frequency/Magnitude chart

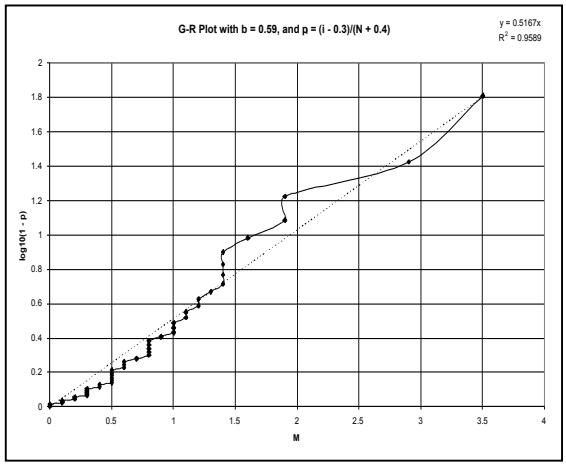

Figure 2(b) : G-R Probability Plot

Simulated G-R catalogues

Using Eqs. 25 and 26, with a = 1.69 and b = 0.59, eleven 131 year catalogues of earthquake events were generated. Figures 2(a) and 2(b) show analysis of one typical

year of synthetic earthquakes using the Gutenberg-Richter frequency/magnitude method, and with a Gutenberg-Richter reduced variate plot.

Linear regression of the data used in Figure 1(a) estimates a to be approximately 1.68, and b to be about 0.54: which agrees with the actual input parameters used to generate the data. Similar linear analysis of the data plot in Figure 2(b) estimates the b parameter to be 0.52.

Visual inspection of Figure 2(b) shows that, although it is reasonable to use the G-R relation to analyse the earthquakes with synthetic magnitudes up to 1.3, events above that magnitude should not be so treated in this particular case.

Figures 3(a) and 3(b) show analysis of the same 131 year of synthetic earthquakes using the Gumbel extreme event method, with the full annual extreme data set, and with the extreme of the annual extreme values truncated.

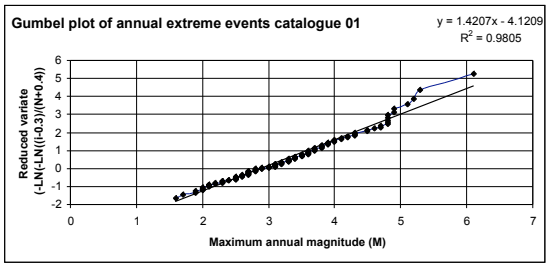

Figure 3(a) : Gumbel analysis full data set

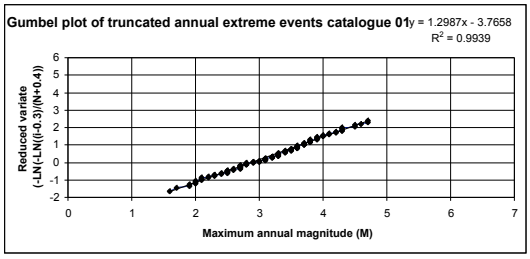

Figure 3(b) : Gumbel analysis truncated data set

Tables 2 and 3 provide a listing of the a and b parameter estimations and averages obtained using the Gumbel extreme value method of analysis, from the full extreme data set, and with the extreme of the extreme values omitted. It can be seen that both methods are capable of recovering the a priori parameter values, and that using the full data set provides the better relative errors.

	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7	Set 8	Set 9	Set 10	Set 11	Avg	Rel Err
a	1.79	1.51	1.67	1.71	1.76	1.70	1.89	1.57	1.61	1.63	1.55	1.67	1.8%
b	0.62	0.52	0.56	0.61	0.64	0.59	0.65	0.56	0.56	0.58	0.55	0.59	1.7%

Table 2: Parameter estimations and average using full extreme data set.

	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7	Set 8	Set 9	Set 10	Set 11	Avg	Rel Err
a	1.64	1.59	1.62	1.70	2.00	1.71	1.82	1.64	1.69	1.59	1.58	1.69	2.4%
b	0.56	0.54	0.55	0.61	0.72	0.60	0.63	0.59	0.58	0.57	0.55	0.59	3.4%

Table 3: Parameter estimation and averages using truncated data set.

Summary

It has been demonstrated that analysis of multiple synthetic earthquake catalogues, derived from a Gumbel seismicity model, using Gumbel distribution plotting of annual extreme earthquake magnitudes, is capable of estimating the a priori a and b parameters values within a relative error of 2%. There is a 95% confidence that individual estimations of α and β will be within 15% and 5% respectively of the true value.

Acceptable parameter estimates are obtained using either full annual extreme data sets, or truncated data sets with the extreme of the extreme values omitted from the data plot, but the full data set provides smaller relative errors.

References

Bury K. 1999, Statistical Distributions in Engineering, Cambridge University Press, ISBN 0 521 63506 3.

Fisher R.A. & Tippett L.H.C. 1928, Limiting forms of the frequency distribution of the largest or smallest member of a sample, Proc. Cambridge Phil. Soc., 24:180.

Gumbel E.J. 1958, Statistics of Extremes, Dover Publications Edition 2004, ISBN 0 486 43604 7, originally published Colombia University Press, 1958.

Johnson N.L., Kotz, S. & Balakrishnan N. 1995, Continuous Univariate Distributions, Volume 2, 2nd Ed., John Wiley & Sons. ISBN 0 471 58494 0.

Lomnitz C. 1974, Global Tectonics and Earthquake Risk, in Developments in Geotectonics 5, Elsevier Scientific Publishing Company.