Bull. Ind. Soc. Earth. Tech., Paper No.368, Vol.34 No.3, Sept. 1997, pp 159- 170

STATISTICAL ANALYSIS OF COMPLETENESS OF SEISMICITY DATA OF THE HIMALAYAS AND ITS EFFECT ON EARTHQUAKE HAZARD DETERMINATION

D. Shanker and M. L. Sharma

Department of Earthquake Engineering University of Roorkee, Roorkee - 247667 INDIA

ABSTRACT

Seismicity data analysis of the Himalayas and its vicinity has been carried out to see magnitude homogeneity and its completeness. In the present study the 270-years sample of earthquakes known to have occurred in the region between 1720 and 1990 is evaluated for completeness along with the question of fitting the frequency formula $\log N = a - bM$ to biased samples that are short with respect to recurrence interval of the largest earthquakes contained in them has been studied. A method proposed by Stepp (1972) has been applied to determine the interval in a magnitude class over which that class is homogeneous. The analysis suggests that the data is complete for the sets $4.5 < M < 5.4$, $5.5 < M < 6.4$, and $6.5 < M < 7.4$ for the past 25 years (1965-1990), 40 years (1950-1990) and 85 years (1905-1990). For magnitudes greater than 7.5 the data is complete for the whole 270 years which may be used to fit the frequency formula. The same complete sets are used to compute b-value and return periods. A b-value of 0.84 is suggested for the whole region.

Key Words: Homogeneity, completeness, recurrence rate, maximum likelihood, Least-square.

INTRODUCTION

A necessary first step in any data analysis is an investigation of available data set to assess the nature and the degree of its completeness. There are several methods of organisation, presentation and reduction of observed data which facilitate its interpretation and evaluation. In the case of analysis of earthquake data, it should be pointed out explicitly that the data used in such studies must be homogeneous and complete. In other words, for homogeneity it must be noted that earthquake data set used for the analysis should be in the same scale. Use of a complete data mean, use of a data sample which includes all earthquakes which occurred in a certain time period

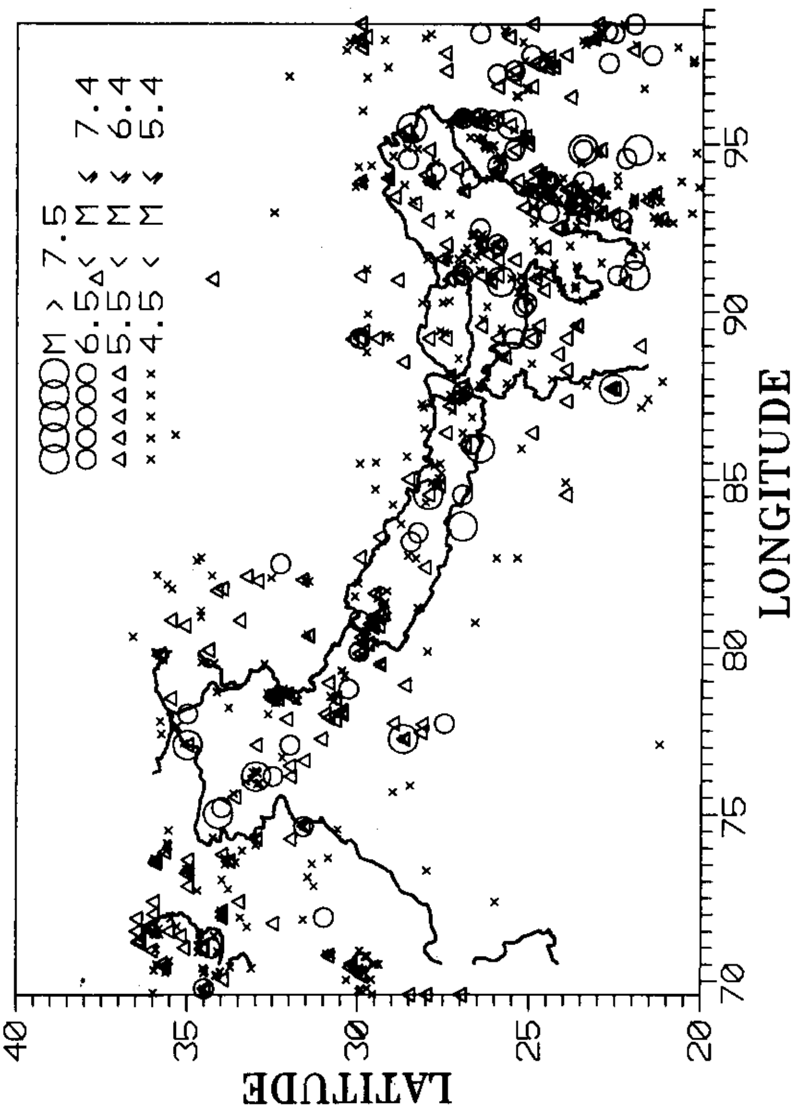

FIG. 1 SEISMICITY OF THE HIMALAYA AND ITS VICINITY FOR THE PERIOD 1720-1990

and have magnitudes larger than a certain minimum value, and is necessary for such kind of work.

The relationship between the size (magnitude, M) and the frequency N (number) of earthquakes over an area during a given period given by Gutenberg and Richter is

$$\log N_k = a_k - bM \tag{1}$$

Where a and b are constants and k is the number of years covered by the data sample. The validity of the empirical recurrence relation (1) has been confirmed by many other scientist for different seismically active regions of the world (Ishimoto and Iida, 1939; Bath, 1981; Tsapanos, 1985; Wang, 1988). It has been suggested that the formula holds for microearthquakes (Mogi, 1962, 67 and Scholz, 1968), showing that it must reflect some fundamental physical property of the fracture process. The relation is also considered to be closely related to tectonic characteristics of a region for earthquake hazard mapping.

In general the incompleteness of existing sample of earthquake data make it difficult to obtain fits of equation (1) that are thought to be represent true long- term recurrence rates. All earthquake catalogues are biased against small shocks, because of Seismograph station density or, the early records. population density. Accordingly, the bias is more severe in successively earlier reporting periods. Southern California catalogue is homogeneous in magnitude 4.0 and larger events since 1933 and in magnitude 3.0 and larger events since 1953 (Knopoff and Gardner, 1969); for the seismogenic region in the Aegean and surrounding area, the data are complete: for Ms > 4. 5 since 1965, for Ms > 5. 5 since 1911, for Ms > 6. 0 since 1901 and for Ms > 6. 5 since 1845 (Papazachos et al, 1997). To fit equation (1) using an earthquake catalogue, the choice must be between using a short sample period that is complete in small events or a longer sample period that is complete in only large events. It has been suggested (Allen et al, 1965) that a 29 year sample drawn from small regions of the dimension of interest in earthquake hazard mapping may not give earthquake recurrence estimates that represent long-range seismicity. Therefore, it is necessary to use longer samples that give more accurate statistical averages of the large earthquakes of primary engineering interest.

The main purpose of this paper is to review and interpret the qualitative and quantitative means of the existing data base for the Himalayas and vicinity (Fig. 1) between 1720 and 1990 and analyse for completeness. In other words, to determine for which time period the data of a given magnitude interval are complete. A procedure described and applied by Stepp (1972) and also by Bollinger (1973) was used and the fitting of the frequency formula that makes use of the complete 270 year sample was determined.

DATA SET USED

The earthquake catalogue from historical period 1720 to 1986 used in the present study are taken from the excellent study of Chandra (1992) and latter periods data (1987-1990) compiled by the first author are taken from the National Geophysical Data Centre, Boulder, Colorado, USA. As the agency uses large number of observation to derive magnitude and epicenter, these estimates may be considered to be reliable. The catalogue before 1900 reports only intensity data. The intensities have been converted to magnitudes ($M = 1 + (\frac{2}{3}) I_o$) given by Gutenberg and Richter (1956). The spatial distribution of earthquakes with magnitude M > 4.5 are shown in Figure 1.

GENERAL ANALYSIS OF SEISMICITY DATA

For this purpose the number of earthquakes per decades was grouped in four magnitude ranges, $4.5 < M < 5.4$; $5.5 < M < 6.4$, $6.5 < M < 7.4$ and $M > 7.5$. Table 1 describes the statistics of earthquakes reported in each decade since the beginning of the available historical record. Figure 2 demonstrate the histogram representation of data listed in Table 1 for the whole catalogue from 1720 through 1990. It is clear from histogram that the most significant jump in the total number of reported events occur in the decade 1980-1990. The whole catalogue shows that period 1720-1800 the data are poorly reported may be due to the lack of observations, and from 1801-1870 the slightly good reporting of data or better recording. The interval 1871-1920 again poor quality of data recording/reporting and from 1921-1990 good reporting of total number of events occur.

In conclusion, the general analysis suggests that considered region shows fluctuations in seismic activity as low-high and low-high, alternatively in a cyclic form with a periodicity of 65 years. In other words, this area may seek bigger earthquake in near future up to 2010.

ANALYSIS OF DATA COMPLETENESS

To analyse the nature of the completeness of the data sample in detail, considered earthquakes are grouped in several magnitude classes and each magnitude class is modeled as a point process in time. Use is made of the estimate of a sample mean which is inversely proportional to the number of observations in the sample (Stepp, 1972). Thus the variance can be made as small as desired by making the number of observation in the sample large enough, provided that reporting is complete in time and the process is stationary i. e., the mean variance and other

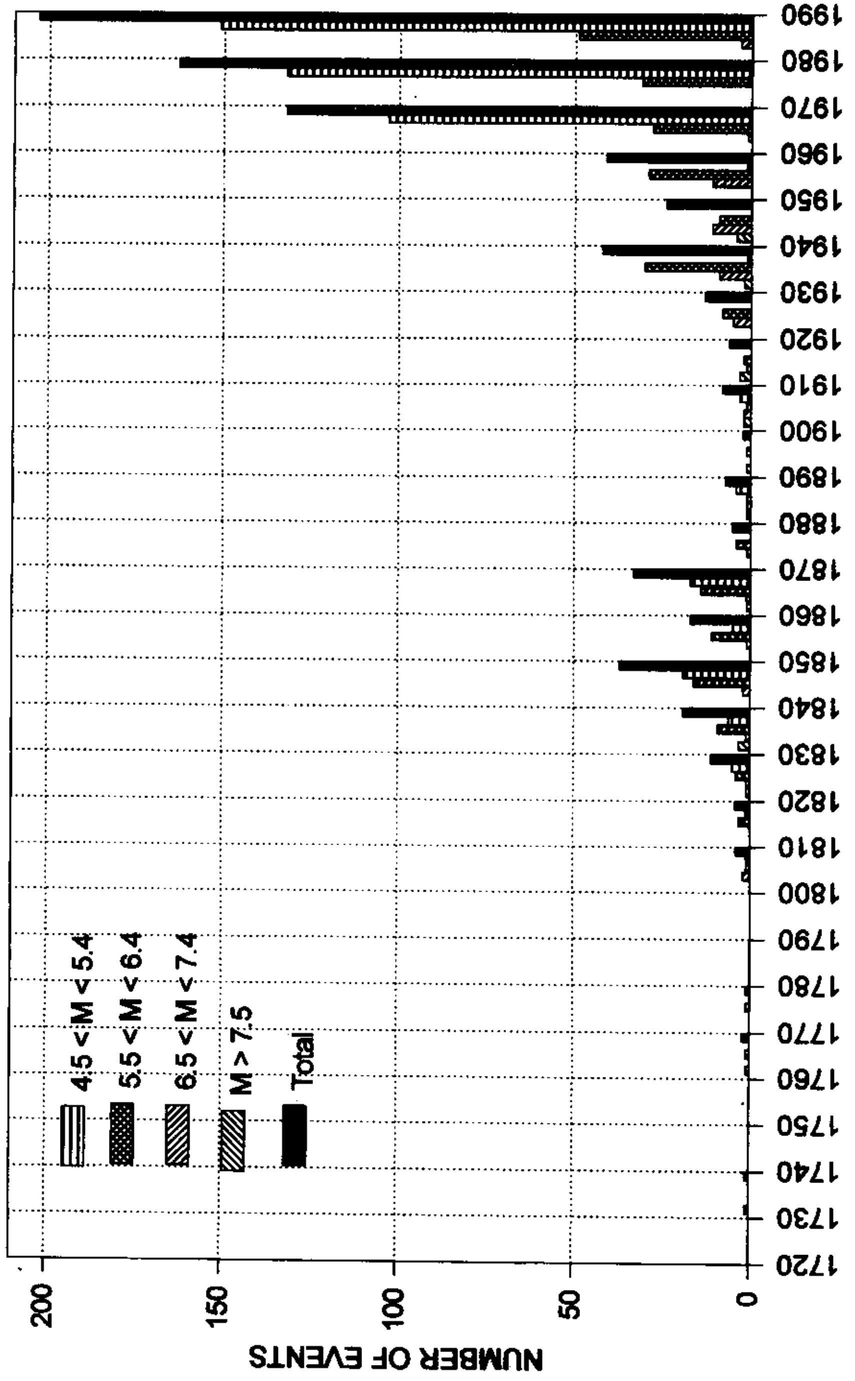

FIG. 2 HISTOGRAM OF THE EVENTS FOR THE CONSIDERED REGION

TABLE I NUMBER OF EARTHQUAKES REPORTED IN EACH DECADE SINCE THE BEGINNING OF THE AVAILABLE HISTORICAL RECORDS

TIME IN YEARS	NUMBER OF EARTHQUAKES				
	$4.5 < m < 5.4$	$5.5 < m < 6.4$	$6.5 < m < 7.4$	$m > 7.5$	TOTAL
1721-1730	0	0	0	0	0
1731-1740	0	0	0	1	1
1741-1750	0	0	0	0	0
1751-1760	0	0	0	0	0
1761-1770	0	1	0	1	2
1771-1780	0	1	0	0	1
1781-1790	0	0	0	0	0
1791-1800	0	0	0	0	0
1801-1810	1	1	2	0	4
1811-1820	1	3	0	0	4
1821-1830	5	6	1	1	13
1831-1840	7	10	1	3	21
1841-1850	22	22	2	0	46
1851-1860	6	12	1	0	19
1861-1870	17	15	1	1	34
1871-1880	1	4	1	0	6
1881-1890	4	1	1	1	7
1891-1900	0	1	0	1	2
1901-1910	3	2	2	1	8
1911-1920	0	2	2	2	6
1921-1930	0	15	3	1	19
1931-1940	5	36	7	2	50
1941-1950	3	10	9	3	25
1951-1960	5	37	3	0	45
1961-1970	125	24	1	0	150
1971-1980	153	26	0	0	179
1981-1990	156	44	3	0	203

moments of each observation stay the same. To obtain an efficient estimate off the variance of the sample mean, it is assumed that the earthquake sequence can be modeled by the Poisson distribution. If $x_1, x_2, x_3 \ldots x_n$ are the number of events per unit time interval, then an unbiased estimate of the mean rate per unit time interval of this sample is (Hamilton, 1964)

$$\lambda = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (2)$$

and its variance is

$$\sigma_\lambda^2 = \frac{\lambda}{n} \qquad (3)$$

where n is the number of unit time intervals. Taking the unit time interval to be one year gives

$$\sigma_{\lambda} = \sqrt{\lambda} \Big/ \sqrt{T} \qquad (4)$$

as the standard deviation of the estimate of the mean, where T is the sample length. Thus, assuming stationary process, one can expect that σ behaves as $1/\sqrt{T}$ in the sub- interval of the 270-year sample, in which the mean rate of occurrence in a magnitude class is constant. In other words, when λ is constant, then the standard deviation (σ_{λ}) varies as $1/\sqrt{T}$ where T is the time interval of the sample. If the mean rate of occurrence is constant we expect stability to occur only in the sub-interval that is long enough to give a good estimate of the mean but short enough that it does not include intervals in which reports are complete.

Table 2 illustrate the rate of occurrence as a function of time interval for different magnitude classes. The rate is given as N/T where N is the cumulative number of earthquakes in the time interval T, for sub-interval of the 270-year sample.

TABLE II EARTHQUAKE DISTRIBUTION BY TIME AND MAGNITUDE

TIME PERIOD	TIME INTERVAL	4.5<M<5.4		5.5<M<6.4		6.5<M<7.4		m>7.5	
		N	N/T	N	N/T	N	N/T	N	N/T
1981-1990	10	156	15.600	44	4.400	3	.300	0	.000
1971-1990	20	295	14.750	68	3.400	3	.150	0	.000
1961-1990	30	404	13.467	90	3.000	4	.133	0	.000
1951-1990	40	409	10.225	123	3.075	7	.175	0	.000
1941-1990	50	412	8.240	132	2.640	16	.320	3	.060
1931-1990	60	415	6.917	164	2.733	21	.350	5	.083
1921-1990	70	415	5.929	174	2.486	24	.343	5	.071
1911-1990	80	415	5.188	176	2.200	26	.325	7	.087
1901-1990	90	418	4.644	178	1.978	28	.311	8	.089
1891-1990	100	418	4.180	179	1.790	28	.280	9	.090
1881-1990	110	422	3.836	180	1.636	29	.264	10	.091
1871-1990	120	422	3.517	184	1.533	30	.250	10	.083
1861-1990	130	439	3.377	198	1.523	31	.238	11	.085
1851-1990	140	444	3.171	209	1.493	32	.229	11	.079
1841-1990	150	463	3.0871	225	1.500	34	.227	11	.073
1831-1990	160	469	2.931	234	1.462	35	.219	14	.087
1821-1990	170	474	2.788	238	1.400	36	.212	15	.088
1811-1990	180	475	2.639	241	1.339	36	.200	15	.083
1801-1990	190	476	2.505	242	1.274	38	.200	15	.079
1791-1990	200	476	2.380	242	1.210	38	.190	15	.075
1781-1990	210	476	2.267	242	1.152	38	.181	15	.071
1771-1990	220	476	2.164	243	1.105	38	.173	15	.068
1761-1990	230	476	2.070	244	1.061	38	.165	16	.070
1751-1990	240	476	1.983	244	1.017	38	.158	16	.067
1741-1990	250	476	1.904	244	.976	38	.152	16	.064
1731-1990	260	476	1.831	244	.938	38	.146	17	.065
1721-1990	270	476	1.763	244	.904	38	.141	17	.063

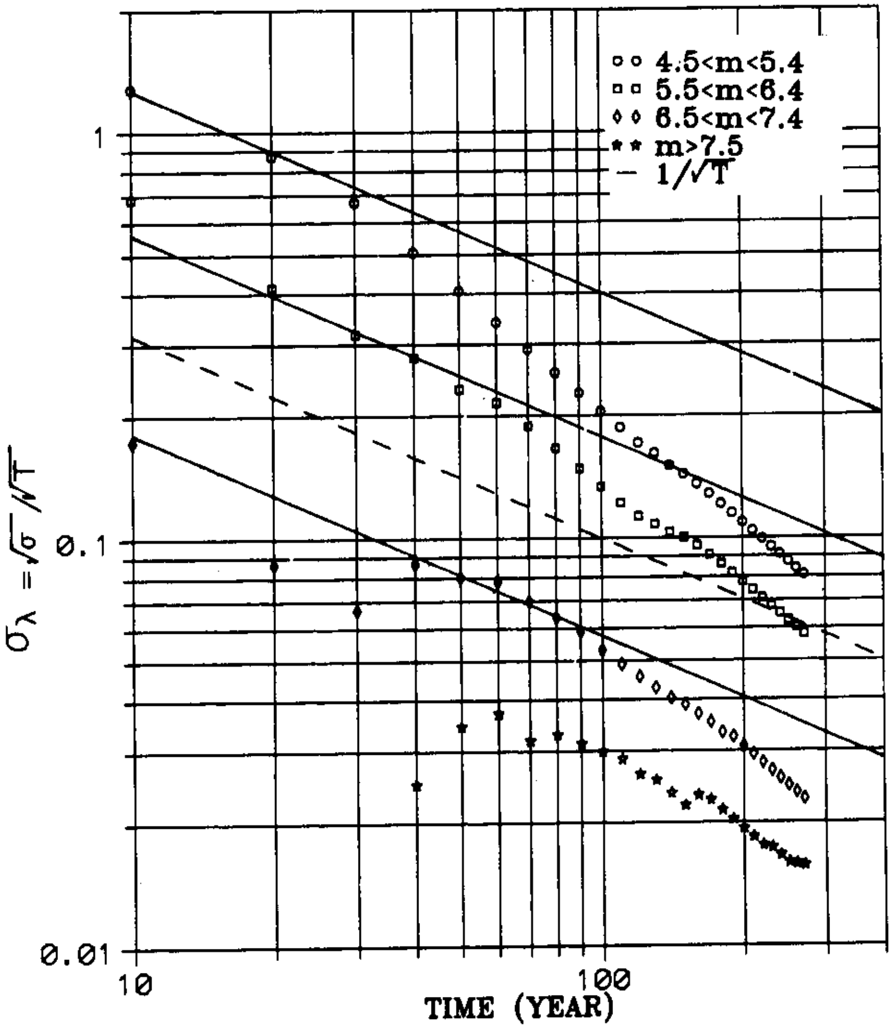

FIG. 3 ESTIMATION OF THE MEAN OF THE ANNUAL NUMBER OF EVENTS AS A FUNCTION OF SAMPLE LENGTH AND MAGNITUDE CLASS

These data are used to compare the standard deviation of the estimate of the mean through equation (4). These results are shown in figure 3. Figure 3 and Table 2 reveal several features significant to statistical treatment of earthquake data, regardless of whether one uses empirical formula $\log N_k = a_k - bM$, the extreme value distribution, or other statistical approaches. For each magnitude interval in Figure 3 the plotted points are supposed to define a straight line relation, as long as the data set for that magnitude interval is complete. For a given seismic region the slope of the lines for all magnitude intervals should be the same. For the sets $4.5 < M < 5.4$ the data appear complete for the past 25 years (1965-1990) and for $5.5 < M < 6.4$ the data seems completely for last 40 years (1950-1990). The slope of the straight line follow the requisite trend upto 1905 for the set $6.5 < M < 7.4$ which shows that the data is completely reported for this magnitude range for the last 85 years. There are less number of reported events in each decades for the magnitude greater than 7.5 compared to the other magnitude intervals. For this data set the plotted points indicate a dearth of earthquakes for the most recent 40 years, but beyond that the data points follow more or less the expected distribution. However, as it is unlikely that any earthquakes of this size were not reported, it is reasonable to assume that this set is complete back to 1721.

From Table 2 and Figure 3 one can conclude that there were 13.53 earthquakes of $4.5 < M < 5.4$ per year (average in the Interval 1965-1990); 3.07 of $5.5 < M < 6.4$ per year (interval 1950-1990); 0. 32 of $6.5 < M < 7.4$ per year (interval 1905-1990) and 0.06 of $M > 7.5$ per year (interval 1720-1990).

TEMPORAL FREQUENCY MAGNITUDE RECURRENCE RELATION

Table 2 demonstrate the number of earthquakes, divided according to magnitude, occurring in 10-year periods of time. An inspection of the Table 2 reveals, that the data set is not complete for the interval 1720 through 1990. The b-slope (b-values) are generally computed for the whole data set without testing the completeness of the data which gives errors in the estimation of b-value. In the present study, the completeness is checked for various magnitude intervals and the same is considered while computing the b-value. The activity of earthquakes per year is taken from the same period for which the data is complete for a particular magnitude range. Table III shows the magnitudes and the time period for which the data is complete. The data in each interval for time period and magnitude range is normalised by respective periods and then the cumulative number of events are taken for computation of b-value. The activity rate is also given in Table III.

Figure 4 shows the plot of the logarithm of the cumulative number of earthquakes per year for $M > m$, where m is the magnitude in particular interval. An interval of 0.5 is taken for grouping the data while computing the b-value. The fitted straight line in least-square sense for complete sets for each magnitude range as follows

$$\log(N) = 5.28 - 0.84M \quad (5)$$

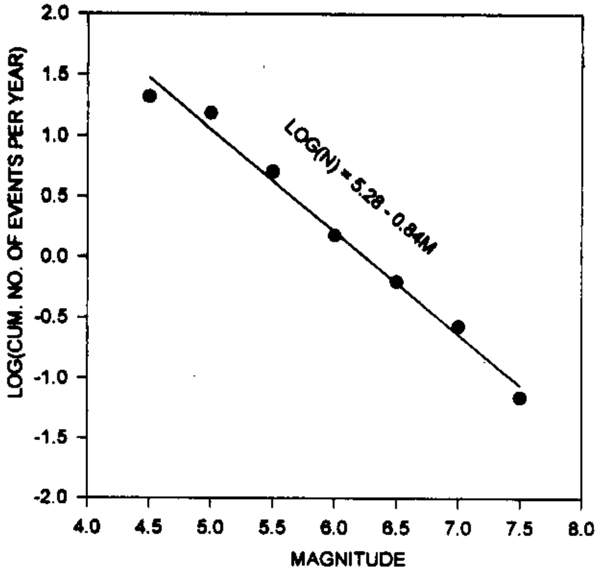

FIG. 4. FREQUENCY MAGNITUDE RELATIONSHIP FOR THE HIMALAYA AND ITS VICINITY.

with correlation coefficient as 0.987. The b-value is also computed for comparison using maximum likelihood method. Application of the maximum likelihood method (Aki, 1965) yield b-value of 0.19 and 0.79 for minimum magnitude considered as 3.10 and 5.0 respectively. Table III also contains recurrence rates for the Himalayas and its vicinity, calculated using equation (5).

TABLE III RECURRENCE RATE FOR THE HIMALAYAS AND ITS VICINITY

M	Completed data in years	Events per year	Return Period
5.0	1965-1990	5.44	0.08
5.5	1965-1990	10.24	0.21
6.0	1965-1990	3.60	0.28
6.5	1950-1990	0.88	0.57
7.0	1950-1990	0.36	3.98
7.5	1905-1990	0.20	10.47

DISCUSSION AND CONCLUSION

It can be advocated from the above analysis that we may create an artificially homogeneous data sample by carefully determining the intervals over which earthquakes in different magnitude classes are completely reported. For each magnitude class the interval must be long enough to establish a stable mean rate of occurrence and short enough that it does not include intervals in which the data are incompletely reported. This amounts, in practice, to minimizing the error of estimate in the mean rate of occurrence of each earthquake class.

Figure 1 demonstrates that the considered region is non uniform seismic region. Therefore one can not conclude that the design earthquake can occur any where in the region. The slope b of the frequency formula describes the distribution of earthquakes in size and is a crucial parameter in earthquake risk calculation. The recurrence period for large magnitude earthquakes is principally of scientific interest, to classify different seismic regions according to their degree of activity. The recurrence relations derived in the present study, are essentially statistical in nature. To justify them on a physical basis one has to assume that the geological processes which result in earthquakes are in a more or less steady- state condition over a length of time. Failure to correct for incomplete reporting in the data sample causes the recurrence rates of large earthquakes to be over estimated, while recurrence rates of small events are under estimated.

ACKNOWLEDGMENT

We acknowledge all the help provided by Prof. S. Basu, Head, Earthquake Engineering during the preparation of the manuscript. The authors sincerely thanks the anonymous reviewer for his helpful suggestions, advice and critical comments that improved the revised version of this manuscript.

REFERENCES

Aki, M. (1965). Maximum likelihood estimate of b in the formula log N = a - b M and its confidence limits. Bull. Earthquake. Res. Inst. Tokyo. Univ. Vol. 43., pp. 237-239.

Allen, C. R., P. S., Amand, C. F. Richter and Nordquist, J. M. (1965). Relation between seismicity and Geological Structure in the southern California region, Bull. Seism. Soc. Am. Vol. 55, pp. 752-797.

Bath, M. (1981). Earthquake Magnitude- Recent Research and Trends. Earth sci. Rev. Vol. 17, pp. 315-398.

Bollinger, G. A. (1973). Seismicity of the Southeastern United States. Bull. Seism. Soc. Am. Vol. 63, No. 5, pp. 1785-1808.

Chandra, U. (1992). Seismotectonics of Himalaya. Current Science, Vol. 62, No. 1-2, pp. 40-71.

Gutenberg, B. and Richter C. F. (1956). Earthquake magnitude. Intensity, energy and acceleration, Bull. Seism. Soc. Am. Vol. 46, 105-145.

Hamilton, W. C. (1964). Statistics in physical Sciences. The Ronald Press Co. Newyork pp. 230.

Ishimoto, M. and K. Iida (1939). Observations sur les seisms euregistre derniermen (I). Bull. Earthq. Res. Inst. Vol. 17, pp. 443-478.

Knopoff, L. and J. K. Gardner (1969). Homogeneous catalogues of earthquakes. Proc. Nat. Acad. Sci., Vol. 63, pp. 1051-1054.

Mogi, K. (1962). Study of elastic shocks caused by the fracture of the heterogeneous materials and its relation to earthquakes phenomena. Bull. Earth. Res. Inst., Vol. 40., pp. 125-173.

Mogi, K. (1967). Earthquakes and fractures. Tectonophysics, Vol. 5, pp. 35-55.

National Geophysical Data Centre, Boulder Collorado, U. S. A. -A catalogue of earthquakes From 1987-1990.

Nuttli, O. W. (1974). Magnitude - Recurrence relation for central Mississippi Valley. earthquakes. Bull. Seism. Soc. Am. Vol. 56., pp. 185-200.

Papazachos, B. C., E. E. Papadimitriou, G. F. Karakaisis and D. G. Panagiopoulos (1997). Long -term earthquake prediction in the Circum- Pacific Convergent Belt. Pure. Appl. Geophys. Vol. 149, pp. 173-217.

Stepp, J. C. (1972). Analysis of the completeness of the earthquake sample in the Puget Sound area and its effect on statistical estimates of earthquake hazard. Proc. Inter. Conf. Microzonation for safer construct. Res. Appl. Seattle, Washington Vol. 64, No. 4., pp. 1189-1207.

Scholz, C. H. (1968). The frequency magnitude relation of micro fracturing in rock and its relation to earthquakes. Bull. Seism. Soc. Am. Vol. 56. ,pp. 185-200.

Tsapanos, T. M., Scordilis, E. M. and Papazachos, B. C. (1988). A World-World Homogeneous and complete Catalogue of earthquakes for the period 1898-1985. . Publ. of Univ. Thessaloniki, Geophysical Laboratory, pp. 188.

Wang, J. H. (1988). b-values of shallow earthquakes in Taiwan. Bull. Seism. Soc. Am. Vol. 78., pp. 1243-1254.