The Application of Coefficient of Variations in Earthquake Forecasting

A.B. Lashak[1*], M. Zare[2], H. Abedi[3], and M.Y. Radan[1]

1. PhD Student, International Institute of Earthquake Engineering and Seismology (IIEES), Tehran, I.R. Iran, *Corresponding Author; email: arefbali@yahoo.co.in
2. Associate Professor, Seismology Research Center, International Institute of Earthquake Engineering and Seismology (IIEES), Tehran, I.R. Iran
3. MSc Graduate of Theoretical Physics, Dept. of Physics, Sharif University of Technology, Tehran, I.R. Iran

ABSTRACT

In this paper it will be investigated that whether it is possible to find some regions in which earthquakes occur as well-behaved random processes (instead of chaotic processes). If so, it will be possible to use analysis methods of random processes in earthquake forecasting. There are two main approaches for earthquake prediction; first, precursory methods based on relationship between abnormal behavior of some geophysical quantity (such as gravitational field, crust conductivity,…) and earthquake occurrence. Second, forecasting methods based on the statistical analysis of earthquakes themselves, which is dealt with in this paper. Each probability distribution function (pdf) in statistics has its own coefficient of variations (CV) which due to it we can have a sense of dispersion and variance level of quantity which obeys that specific pdf and also its future variances. In the case of earthquake occurrence also it is possible to calculate the CV of inter-occurrence times of sequential earthquakes in a specified region and specified time interval, in order to find appropriate subregions in which random processes analysis tools can be used for forecasting future seismic behaviors. Here this idea has been applied to Iran.

Keywords:
Coefficient of variations;
Random processes;
Exponential distribution;
Seismological provinces;
Earthquake forecasting

1. Introduction

Coefficient of variations (*CV*) is one of the useful quantities in descriptive statistics which is defined as division of standard deviation σ by the mean μ of some statistical data [1-2]:

$$C_v = \frac{\sigma}{\mu} \quad (1)$$

The *CV* could be interpreted as a normalized measure of dispersion. It is used to calculate the intensity level of variations and dispersion of data sets. It works better than ordinary standard deviation because it is divided by the mean. For example, the standard deviation of the two numbers 0 and 1 is equal to 0.5 and it exactly satisfies for the two numbers 1000000 and 1000001. It should be noted that the variation of some quantity from zero to one is extremely different to variation from 1000000 to 1000001; because in the first case the quantity is multiplied by infinity and in the second by 1.000001, therefore the dispersion and variance in this case is extremely less and the data are more ordered. This fact will be more illustrated when the *CV* is calculated instead of the standard deviation; the *CV* of 0 and 1 is equal to 1 and the *CV* of 1000000 and 1000001 is equal to 0.0000005 which shows that the variation is much less in the second case.

The *CV* is a tool for measuring the variation rate of statistical quantities. For example in financial fields, the reliability theory states that investment on stocks or goods which have a high *CV* on their daily prices is risky. In probability theory, statistical distributions are classified in two classes, dependent on their *CV* amount: low-variance and high-variance. For exponential distribution, which is often used to model the time between independent events that happen at an average rate λ^{-1} and have relation in

the form of $\lambda e^{-\lambda x}$ as shown in Figure (1), the standard deviation is equal to its mean, therefore its *CV* is equal to one. Distributions with *CV* less than one (such as an Erlang distribution) are considered as low-variance, while those with *CV* greater than one (such as hyper-exponential and power-law distributions) are considered as high-variance.

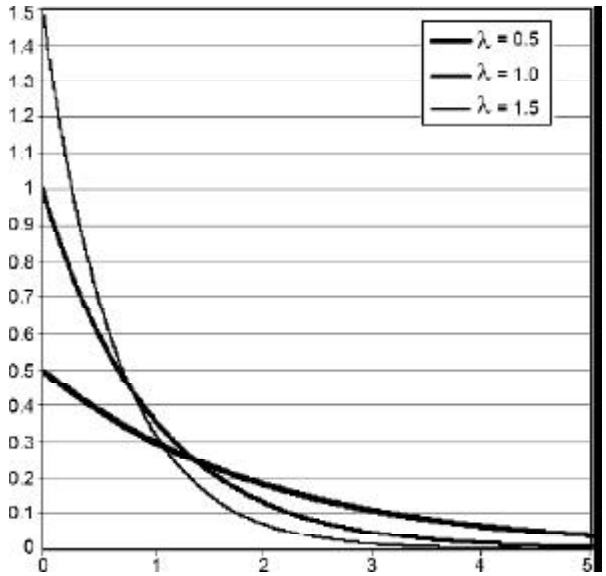

Figure 1. Exponential probability distribution function for various amounts of λ.

As mentioned before, random processes occur in a way which the frequency distribution of their inter-occurrence times of sequential events is exponential. On the other hand, if the distribution of inter-occurrence time of some sequential events is exponential, the number of occurred events in a fixed time interval will have the Poisson distribution. Thus, the random Poisson occurrence [3] refers to well behavior stochastic processes with Poisson distribution for the rate of occurrence of events and this should be distinguished from inter-occurrence time distribution which is exponential.

One of the most important features of *CV* is that it is an unit-less quantity, so it provides the opportunity of comparing between data sets with different units (like the height and weight of some persons). Furthermore, if we calculate the *CV* of a data set, (for example by dimension of time) a constant number for the *CV* will be gained, independent on the unit of numbers (day, hour, minute, second, etc). If the *CV* of a set of numbers is equal to zero, it can be concluded that the numbers are equal and there is no variation through them.

The restriction in usage of *CV* is whenever the mean is equal to zero. Now consider the 20 numbers below; every k numbers $(2 \le k \le 20)$ from the left hand have the *CV* equal to one:

0, 1, 3.8, 7.5, 15.8, 25.8, 39, 55, 75, 97, 124, 155, 190, 228, 271, 319, 370, 427, 489, 555

It can be assumed that these numbers are the time distances between sequential earthquakes in some regions in units such as day. The purpose of representing these numbers is to give the reader an intuition about coefficient of variations (*CV*) equal to one.

In continuation, we would like to study the sequences of random events occurring in time. Suppose starting from a time point $t_0 = 0$, we begin to count the number of events. Then for each time value t, the number of events $N(t)$ that have occurred in time interval $[0, t]$ are obtained. For example $N(t)$ is the number of earthquakes occurred in time interval $[0, t]$, or the number of accidents in a particular crossroad and so on.

Clearly $N(t)$ is a discrete random variable with possible values from $\{0, 1, 2,...\}$. To study the distribution of $N(t)$ the following assumptions are made:

1. All $n \ge 0$, and for any two equal time intervals Δt_1 and Δt_2 the probability of n events in Δt_1 is equal to probability of n events in Δt_2.
2. For all $n \ge 0$, and for any interval $(t, t + s)$, the probability of n events in $(t, t + s)$ is independent of how many events have occurred earlier or how they have occurred. More formally, let $0 \le t_1 < t_2 < t_3 < ... < t_k$ be the given times and $A_i, 1 \le i \le k\text{-}1$ be the event that n_i events occurred in time interval $[t_i, t_i + 1)$. The independent increments mean that $\{A_1, A_2, ..., A_{k-1}\}$ is an independent set of events.
3. The occurrence of two or more events in a very small time interval is practically impossible. Let $N(t)$ be the number of events occurred during $[0, t]$, then

$$\lim_{h->0} \frac{P(N(h) > 1)}{h} = 0 \qquad (2)$$

In other words as $h \to 0$, the probability of two or more events, $P(N(h) > 1)$ approaches zero faster than h does.

By the first condition the random variables $N(t_1)$ - $N(t_2)$ and $N(t_1 + s)$ - $N(t_2 + s)$ have the same probability mass functions, i.e. the probability of

occurrence of n events in the time interval $[t_1, t_2]$ is the function of t_2 - t_1 and not of t_1 and t_2 independently. Properties 1 and 3 result in the following fact that the simultaneous occurrence of two or more events is impossible, i.e. events occur one at a time [3].

Suppose that events occur in time in a way that satisfy the three above conditions, then if for any interval of length $t > 0$, $P(N(t) = 0) = 0$, we will have at least one event for any interval of length t and it can be shown that in this case in any interval of arbitrary length at least one event occur with probability 1. Similarly if $P(N(t) = 0) = 1$ then in any interval of length t no event will occur and in this case any interval of arbitrary length will have no events with probability 1. To avoid these cases, it is assumed:

$$0 < P(N(t) = 0) < 1 \tag{3}$$

If random events occur in time and the three conditions above are all satisfied, $N(0) = 0$ and for all $t > 0$, $0 < P(N(t) = 0) < 1$, then there exists a positive number λ such as:

$$P(N(t) = n) = \frac{(\lambda t)^n e^{-\lambda t}}{n!} \tag{4}$$

The meaning of the above statement is that for all $t > 0$, $N(t)$ is a Poisson random variable with parameter λt. Hence $E[N(t)] = \lambda t$ and $\lambda = E[N(1)]$. It should be noted that the only unknown parameter λ is equal to the expected number of events over a unit time period. This is a very useful equality which can be used to estimate λ in practice.

If the number of events $N(t)$ occurring during a fixed time interval of length t has a Poisson distribution with parameter λt then the corresponding process is called a Poisson process and λ is the rate of the process [4]. Poisson processes are often denoted by:

$$\{N(t) \mid t \geq 0\} \tag{5}$$

Let $\{N(t) \mid t \geq 0\}$ be a Poisson process. Let X_1 be the time of the first event, X_2 the time elapsed between first and second events, X_3 the time between second and third and so on. The sequence of continuous random variables $\{X_1, X_2, ...\}$ is the sequence of interval times of the Poisson process. Let $\lambda = E[N(1)]$, then:

$$P(X_1 > t) = P(N(t) = 0) = e^{-\lambda t} \tag{6}$$

$$P(X_1 \leq t) = 1 - P(X_1 > t) = 1 - e^{-\lambda t} \tag{7}$$

It can be shown that in the case of a Poisson process, as a consequence of the three assumptions, the random variables in the sequence $\{X_1, X_2, ...\}$ are identically distributed. Therefore, for all $n \geq 1$:

$$P(X_n \leq t) = P(X_1 \leq t) = \begin{cases} 1 - e^{-\lambda t} & t \geq 0 \\ 0 & t < 0 \end{cases} \tag{8}$$

F is called exponential distribution if for some $\lambda > 0$:

$$F(t;\lambda) = \begin{cases} 1 - e^{-\lambda t} & t \geq 0 \\ 0 & t < 0 \end{cases} \tag{9}$$

$F(t;\lambda)$ is the cumulative distribution function for X_n, n $\geq$ 1.

It is easy to see that $F(t;\lambda)$ is a distribution function since the corresponding probability density function

$$F(t;\lambda) = F'(t;\lambda) = \begin{cases} \lambda e^{-\lambda t} & t \geq 0 \\ 0 & t < 0 \end{cases} \tag{10}$$

is always non-negative and

$$\int_0^{\infty} \lambda e^{-\lambda t} dt = \lim_{b \to \infty} \int_0^b \lambda e^{-\lambda t} dt = \lim_{b \to \infty} [-e^{-\lambda t}]_0^b = \lim_{b \to \infty} (-e^{-\lambda b} + 1) = 1 \tag{11}$$

A continuous random variable X is called exponential with parameter $\lambda > 0$ if its probability density function is:

$$F(t;\lambda) = \begin{cases} \lambda e^{-\lambda t} & t \geq 0 \\ 0 & t < 0 \end{cases} \tag{12}$$

2. The CV of Earthquake Occurrence Rate in Iran

In order to assess the *CV* of earthquake occurrence rate in Iran, the country was divided into a grid of one in one degree cells and the occurred earthquakes during 1976 to 2008 from *USGS* website [5] was extracted for each cell. Then the inter-occurrence time between sequential earthquakes for each cell was calculated and in this way our primary data was gained. Then, using this data the *CV* of earthquake occurrence rate for each cell was calculated. It should be mentioned that this calculation is repeated for threshold magnitudes 3, 3.5, 4, 4.3, 4.5, 4.6, 4.7, 4.8, 5 and 5.5 and those cells which contained less than 5 earthquakes were considered empty.

We especially concentrated on regions which had *CV* equal to one because the frequency distribution of the data in these regions was exponential. Therefore, it can be concluded that the earthquakes in these regions have occurred as independent stochastic events and are not considered as chaotic and we can take advantage of analyzing tools of stochastic phenomena. For example, in some region with *CV* equal to one, using the average occurrence rate of last earthquakes, it is possible to determine the occurrence time interval of the next event by confidence level of 95%.

The results from regional calculations of *CV* are shown by contours which in all of them the regions without enough data are represented by white color and the regions with *CV* around one (between 0.95 and 1.05) are considered as our target.

In Figures (2) to (5) which are related respectively to the earthquakes greater than 3, 3.5, 4 and 4.3, the overall behavior of contours are the same and some area other than Iran are included. In Figures (6) to (9) which are related respectively to the earthquakes greater than 4.5, 4.6, 4.7 and 4.8, it is obvious that the regions without enough data have increased and the overall *CV* have approached to one (See the scale column).

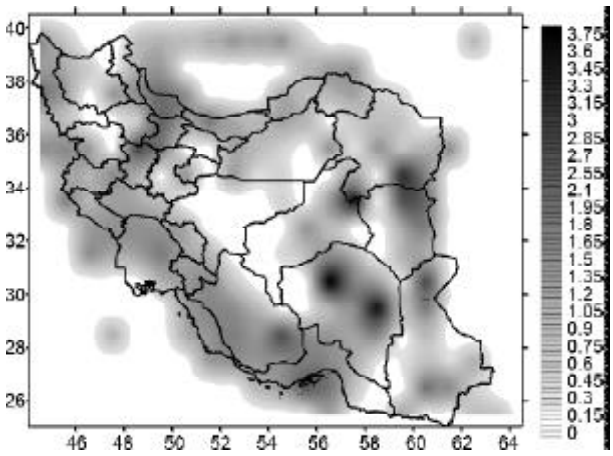

Figure 2. Earthquakes greater than 3.

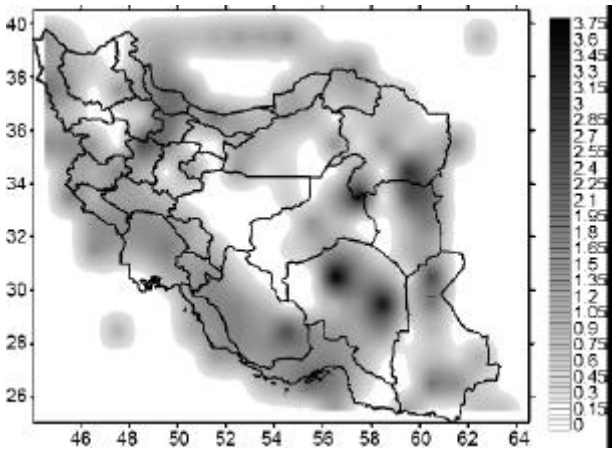

Figure 3. Earthquakes greater than 3.5.

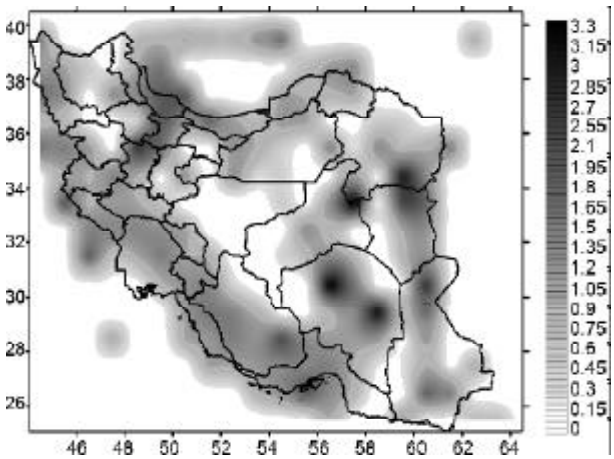

Figure 4. Earthquakes greater than 4.

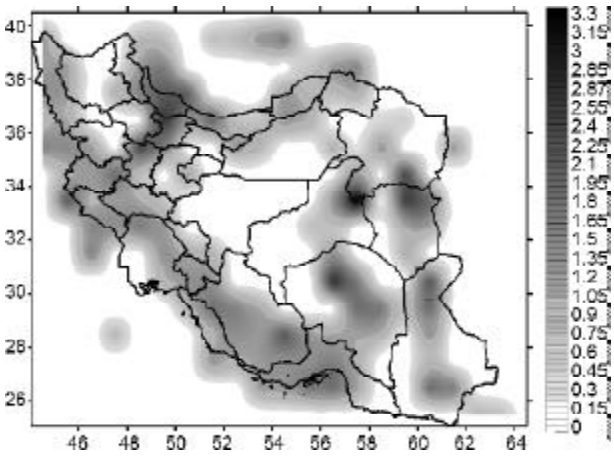

Figure 5. Earthquakes greater than 4.3.

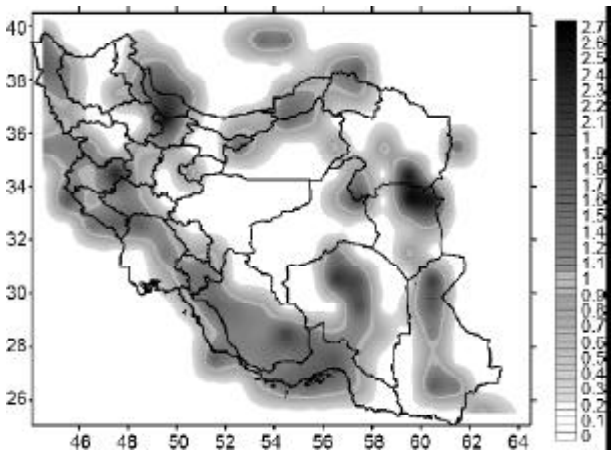

Figure 6. Earthquakes greater than 4.5.

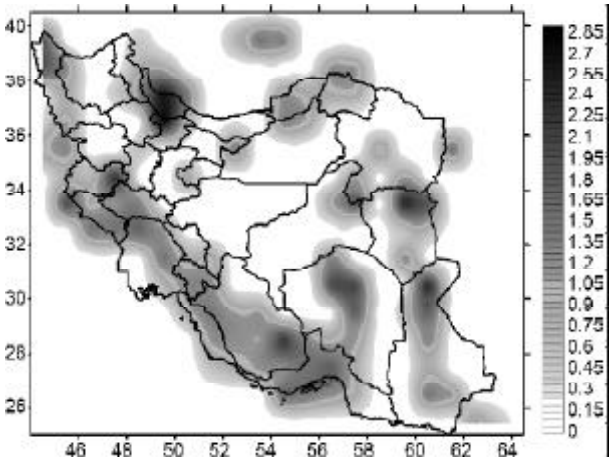

Figure 7. Earthquakes greater than 4.6.

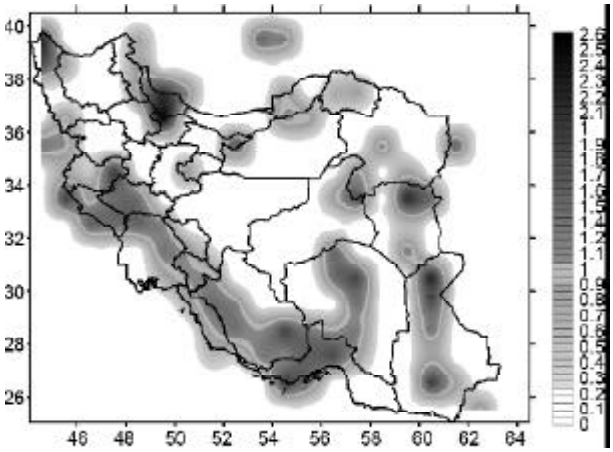

Figure 8. Earthquakes greater than 4.7.

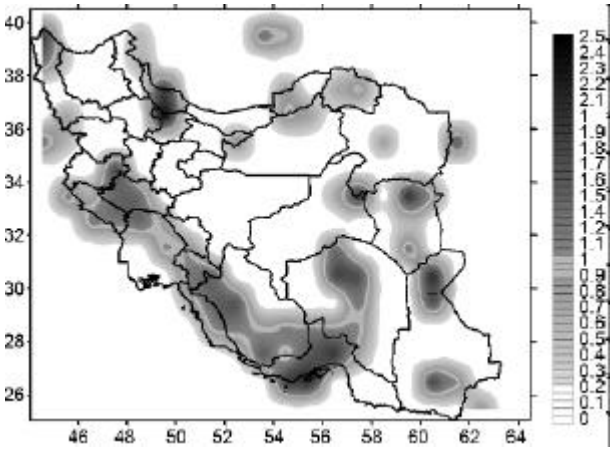

Figure 9. Earthquakes greater than 4.8.

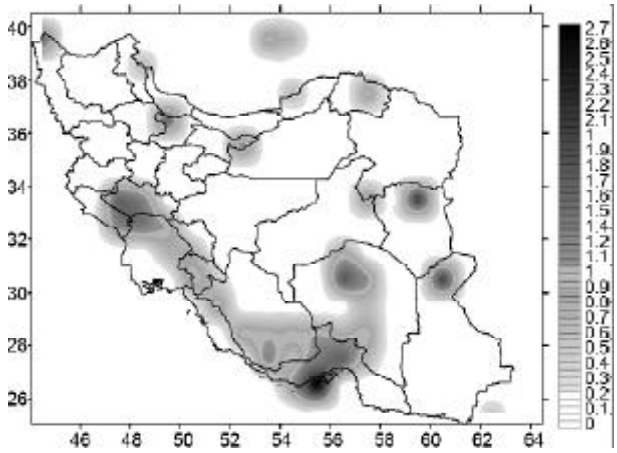

Figure 10. Earthquakes greater than 5.

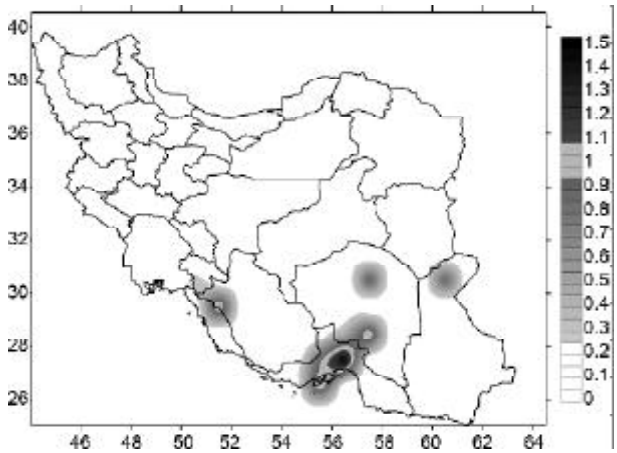

Figure 11. Earthquakes greater than 5.5.

For earthquakes greater than 5 and 5.5, as shown in Figures (10) and (11) the regions free of data have not been specified and in conclusion they were neglected due to the lack of data.

Therefore, as the above figures propose, we selected $M = 4.5$ as the threshold magnitude and performed our classification based on the *CV* amounts in different regions using a magnitude edited earthquake catalogue, see Figure (12).

In order to unify the different units for earthquake magnitude and to increase the accuracy of selecting earthquakes, a magnitude convertor formula [6] was applied to *NEIC* earthquake catalogue and then the earthquakes greater than 4.5 were selected, see Table (1).

According to Figure (12) there is a region within latitude 35-38*N* and longitude 53-56*E* which its *CV* of earthquake occurrence rate is around one and less than it. Therefore, it is expected that events in

Table 1. Magnitude convertor formula.

Magnitude Scale	Relationship
Ml = Mw	-
Mw = 0.99 × Ms + 0.08	6.1 < Ms ≤ 8
Mw = 0.67 × Ms + 2.7	3 ≤ Ms ≤ 6.1
Mw = 0.85 × Mb + 1.03	3.5 ≤ Mb ≤ 6.2

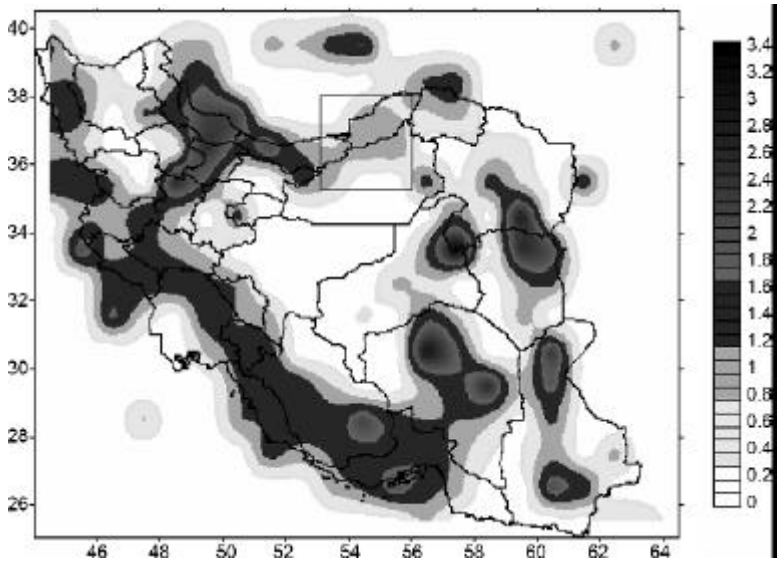

Figure 12. Classification of seismogenic regions based on the CV amounts for earthquakes greater than 4.5.

this area follow the stochastic processes pattern. To investigate this idea, the frequency distribution of the data should be calculated. The data is the inter-occurrence times of sequential earthquakes in the region from 1976 to 2008. In this temporal and spatial interval, 81 earthquakes greater than 4.5 has been registered, therefore there were 80 time differences as in our data. These data have been shown in Table (2).

The numbers in Table (2) are time differences between the occurrence of sequential earthquakes in unit of day. Their *CV* is independent of time unit (as mentioned before) and is equal to 1.133 and their average is equal to 146.873. There is an interesting point about the constancy of average amount, from stochastic phenomenology point of view; the average of each 30-40 sequential numbers is almost equal to the total average and it is confirmed that seismicity regime in this region is well behaved and stable.

The other thing to be checked in order to make sure about the stochastic characteristic of seismicity regime in this area was the frequency distribution of the data. The frequency distribution of these data according to Figure (13) coincided the exponential distribution curve with an acceptable accuracy.

The next step was to examine the predictability of future earthquakes in this region by using the rules and relations of the probability theory. It can be shown that for the stochastic processes with temporal average β, the probability that the next event does not take place in time interval t from the last one is:

Table 2. Inter-occurrence times of our data.

80 Temporal Intervals between Sequential Earthquakes				
108.8684	35.36903	192.8134	488.5148	175.8917
226.6734	387.9253	209.907	452.1418	18.77651
124.2828	44.92582	86.291	189.0771	343.1318
66.55239	67.91525	15.25433	21.47965	16.18153
113.0151	0.019744	39.62213	47.78054	2.952722
463.2528	0.004979	38.64066	0.197112	194.8437
86.34789	0.023449	289.2736	82.46425	0.065842
29.68835	8.22557	282.9923	412.235	0.600513
100.9613	17.89911	11.25255	150.6276	24.2739
472.144	121.9135	9.00033	100.3795	69.93556
103.6654	657.6103	3.636673	6.990974	322.465
2.930196	116.0371	245.3079	13.74526	385.9607
149.8042	493.0078	80.53658	206.2831	50.80142
126.4517	128.8011	37.42965	44.32275	403.1543
495.3139	27.15274	0.135169	681.2739	120.5
231.3685	64.66862	11.28886	80.31841	16.28322

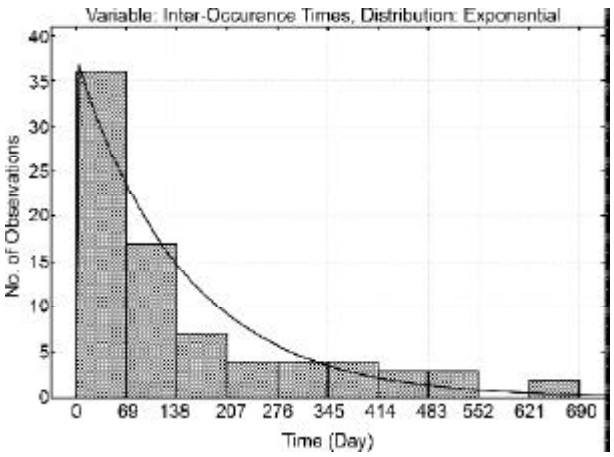

Figure 13. Frequency distribution of inter-occurrence times.

$$P' = exp(-t/\lambda) \quad (13)$$

Therefore,

$$P = 1 - P' = 1 - exp(-t/\lambda) \quad (14)$$

is the probability of next event occurrence in time interval t from the last event. Solving it with respect to t:

$$t(P) = \lambda \ ln(1/(1-P)) \quad (15)$$

Therefore, $t(P)$ is the time interval from the last event in which the next event occurs with probability P.

Suppose that having the information pertaining to 33 earthquakes and therefore having 32 data, the aim was to determine how long should be elapsed from the last earthquake occurrence until 34^{th} earthquake occurs. To do so, the average of 32 data should be replaced instead of λ and 0.7 instead of P in Eq. (15). In this way $t(P)$ was calculated equal to 190.861. Referring to Table (2), it is seen that the 34^{th} earthquake occurs 192.813 days after 33^{th} earthquake i.e. by two days difference with predicted time interval.

In order to perform a more comprehensive assessment of the success level of the above calculation, we repeated it for probabilities 0.6, 0.7, 0.8 and 0.9 to predict the time intervals number 32 to the end. The first 31 data was skipped because much data is needed to perform statistical calculations; therefore the calculations began from 1990 i.e. 32 data which is the temporal distance between earthquakes 32 and 33. The result is shown in Table (3).

In this table, column N is the number of data, column *Data* is the amount of real data occurred in the past (temporal distances between occurred earthquakes) and the left hand columns are predicted time intervals (using formula 15) for occurrence of $N+1^{th}$ earthquake with probability 0.6, 0.7, 0.8 and 0.9, respectively. These numbers should be compared with their own corresponding data (N^{th} data).

Whenever predicted time interval is smaller than real data in each row, that cell has been shown by gray. For example in column 0.6, there are 18 gray cells and 32 white cells among 50 cells, hence 64% of events have occurred in predicted time intervals with 0.6 probability and this is an acceptable compatibility. These numbers for column 0.7, 0.8 and 0.9 are 68%, 78% and 86%, respectively. In conclusion it can be said that the temporal seismicity regime in this region obeys well-behavior stochastic processes pattern available in statistics and mathematics.

Table 3. Calculations resulted from formula (15).

N	Data	P = 0.6	P = 0.7	P = 0.8	P = 0.9
32	27.15274	152.13475	199.89955	267.22025	382.30575
33	64.66862	148.02975	194.50573	260.00994	371.99013
34	192.81345	**145.2556**	**190.8605**	255.13713	365.01872
35	209.90699	**146.2076**	**192.1115**	256.80938	367.41116
36	86.291	147.56431	193.89416	259.19241	370.8205
37	15.25433	145.60727	191.32267	255.75491	365.90255
38	39.62213	141.95088	186.51831	249.33258	356.71427
39	38.64066	139.0956	182.76657	244.31735	349.53911
40	289.2736	**136.3669**	**179.1812**	**239.5245**	342.68212
41	282.99232	**139.6667**	**183.517**	**245.3205**	350.9743
42	11.25255	142.65763	187.44695	250.57396	358.49029
43	9.00033	139.42965	183.20551	244.90411	350.37857
44	3.63667	136.30625	179.10148	239.41795	342.52965
45	245.3079	**133.2138**	**175.0381**	**233.9862**	334.75859
46	80.53658	135.29473	177.77237	237.64125	339.98776
47	37.42965	133.92807	175.97663	235.24074	336.55342
48	0.13517	131.76217	173.13071	231.4364	331.11063
49	11.28886	128.96135	169.45054	226.51685	324.07234
50	488.51482	**126.4902**	**166.2035**	**222.1763**	**317.8624**
51	452.14185	**133.0439**	**174.8148**	**233.6876**	**334.3314**
52	189.07708	**138.6688**	**182.2058**	243.56778	348.46671
53	21.47965	139.3469	183.09678	244.75876	350.17062
54	47.78054	137.04564	180.07301	240.71667	344.3877
55	0.19711	135.28593	177.76081	237.62579	339.96565
56	82.46425	132.78398	174.47334	233.23119	333.67839
57	412.23498	**131.7436**	**173.1063**	**231.4037**	**331.0639**
58	150.6276	**136.1361**	178.87794	239.11914	342.10215
59	100.37945	136.16916	178.92134	239.17715	342.18514
60	6.99097	135.40722	177.92018	237.83883	340.27044
61	13.74526	133.22075	175.04724	233.99837	334.77598
62	206.28312	**131.2103**	**172.4056**	230.46709	329.72387
63	44.32275	132.15794	173.65075	232.13157	332.10519
64	681.27394	**130.6814**	**171.7106**	**229.5381**	**328.3947**
65	80.31841	138.51575	182.00467	243.29886	348.08198
66	175.89174	**137.5014**	180.6718	241.51712	345.53289
67	18.77651	137.86546	181.15022	242.15666	346.44785
68	343.13178	**136.0373**	**178.748**	**238.9455**	**341.8537**
69	16.18153	138.69953	182.24615	243.62167	348.54381
70	2.95272	136.87787	179.85256	240.42198	343.96609
71	194.84372	**134.9333**	**177.2975**	237.00648	339.07961
72	0.06584	135.5562	178.11594	238.10051	340.64482
73	0.60051	133.64781	175.60838	234.74848	335.84914
74	24.2739	131.79923	173.17942	231.5015	331.20378
75	69.93556	130.29845	171.20744	228.86542	327.43239
76	322.46504	**129.4036**	**170.0317**	**227.2937**	325.18375
77	385.96072	**131.6179**	**172.9411**	**231.1829**	**330.748**
78	50.80142	134.53937	176.77985	236.31447	338.08958
79	403.1543	**133.3966**	**175.2783**	**234.3073**	**335.218**
80	120.5	136.4224	179.25409	239.62197	342.82153
81	16.28322	136.09317	178.82149	239.04368	341.99419

3. Results and Suggestions

The following results are obtained in this study:

1. A new pattern has been presented to identify the identical regions with the same seismicity rate in Iran. In this pattern Iran is classified into four different seismic provinces based on the amount of *CV* in each province:
   - Seismicity regime *A* with *CV* in interval 0.2-0.8;
   - Seismicity regime *B* with *CV* in interval 0.81-1.2;
   - Seismicity regime *C* with *CV* in interval 1.21-1.7; and
   - Seismicity regime *D* with *CV* greater than 1.71.
2. In order to forecast the future earthquakes greater than the threshold magnitude:
   - Periodic and ordered patterns can be used in regions with *CV* less than 0.2;
   - Low-variance distributions can be used in region *A*;
   - Exponential distribution can be used in region *B*;
   - High-variance and power-law distributions and patterns related to clustering in region *C*;
   - Fractal analysis methods or chaos analysis methods in necessary occasions (very great *CVs*) in region *D* [7].

   It should be notified that if the area of investigation consists of combination of regions, the most severe method should be selected for analyzing, for example in region 54-56*E*, 26-28*N* fractal analysis should be applied.
3. It is possible to develop this work in order to determine the threshold magnitudes corresponding to *CV*=1 in different regions in Iran in which the events can be considered as well defined stochastic events and also to take advantage of exponential distribution for forecasting the occurrence time of the next similar earthquakes [3].

References

1. Hoel, G.P. (1984). "Introduction to Mathematical Statistics", 5th Edition, Whiley Publications, ISBN: 0471890456.
2. Metcalfe, A.V. (1994). "Statistics in Engineering: A Practical Approach, 1st Edition, Chapman and Hall/CRC Press, ISBN: 0412492202.
3. Kagan, Y.Y. and Jackson, D.D. (1991). "Long-term Earthquake Clustering", *Geophysical Journal International*, **104**(1), 117-133.
4. Devore, J.L. (2007). "Probability and Statistics for Engineering and the Sciences", Thomson Books, ISBN: 0495382175.
5. USGS Website: http://heic.usgs.gov.
6. Scordilis, E.M. (2006). "Empirical Global Relations for *MS*, *Mb*, *ML* and Moment Magnitude, *Journal of Seismology*, doi: 10.1007/S10950-006-9012-4.
7. Kagan, Y.Y. and Jackson, D.D. (2000). "Probabilistic Forecasting of Earthquakes", *Geophysical Journal International*.