This is an extended version of the paper presented in SEE7 conference, peer-reviewed again and approved by the JSEE editorial board.

Application of Copula Theory to Develop Techniques for Earthquakes Forecasting

Mostafa Allamehzadeh[1*], Mohammad Kavei[2] and Mehrdad Mostafazadeh[1]

1. Assistant Professor, Seismological Research Center, International Institute of Earthquake Engineering and Seismology (IIEES), Tehran, Iran,
* Corresponding Author; email: zadeh66@hotmail.com
2. Assistant Professor, Department of Physics, University of Hormozgan, Bandar Abbas, Iran

Received: 24/08/2015
Accepted: 09/03/2016

ABSTRACT

Recent advances made in forecasting Earthquakes using clustering analysis techniques are being run by numerical simulations. In this paper, the Gaussian Copula clustering technique is used to obtain Earthquake patterns such as the Doughnut Earthquake pattern to better predict medium and large events. Copulas methods can involve recognizing precursory seismic patterns before a large earthquake within a specific region occurs. The observed data represent seismic activities situated around IRAN in the 1980-2014 time intervals. This technique is based on applying cluster analysis of earthquake patterns to observe and synthetic seismic catalog. Earthquakes are first classified into different clusters, and then, patterns are discovered before large earthquakes via Copulas simulation. The results of the experiments show that recognition rates achieved within this system are much higher than those achieved only during the feature map is used on the seismic silence and the Doughnut pattern before large earthquakes

Keywords:
Pattern recognition; Copula methods; Clustering; Earthquake forecasting

1. Introduction

Natural incidents, such as earthquakes occur in most places of the world and cause incalculable human losses, in addition to billions of dollars in damages each year.

For example, 622,000 people died in natural disasters in a ten-year period from 1992. It is reported that more than half of the fifty most significant of such disasters were recorded in 2001, while earthquakes, caused the highest economic losses. In these situations, historical records containing observations from the past are usually the only source of information.

Predicting earthquakes, in time and space, have not been sufficiently reliable. Some of the reasons are the varied trigger mechanisms and insufficient instrumentation. Field et al. [1] classify two types of credible earthquake predictions in California, which has a recent 180-year record. One is a general forecasting method, which gives probabilities of occurrences over a long period of time. The second attempts to be specific by stating the time interval, region and range of magnitude. The Bulletin of the Seismographic Stations of the University of California lists 3638 earthquakes of magnitude, in the range of $3\_0 \leq ML \leq 7\_0$, observed during the period 1949 to 1983 over an area of 280,000 square kilometers in northern and central California.

We consider the study of the statistics of extreme earthquakes to be a first step in the mitigation of these national disasters. Allamehzadeh and Mostafazadeh [2] present an overview of Copula methods in the context of earthquake forecasting.

Several surveys of Copula theory and applications have appeared in the literature to date: Nelsen [3] and Joe [4] are two key textbooks on Copula theory, providing clear and detailed introductions to Copulas and dependence modeling, with an emphasis on statistical foundations.

Details of seismic patterns are controlled by tectonic environment (geometry faults and strain rate) and by the inhomogeneous of the fault plane. In order to explain the seismic clusters, a simple Asperity model is hypothesized. In this model, fault system possesses an asperity that has the number of under-thrust faults, and by increasing the tectonic stresses of under-thrust faults located in the weak area accidently, the pre-seismic earthquake will be active until the fault plane is calm down from the side of earthquake. In the event that tectonic stress would increase, the Asperity breaks and arrangements for essential tremor will occur along with the entire fault zone. The driving force distribution acting as spirit earthquakes can happen or spatial-temporal variations in the stress does not appear, likely due to the change in the pattern of seismic fault plane imposes. Because of an earthquake to the others, seismicity patterns are another reason why they cannot be used alone in earthquake prediction and measurement of physical parameters. According to Jones and Molnar [5], about 44% of shallow earthquakes in the world since the seismicity rate increase occurred in different spatial scales. This evidence indicates that seismic activity before major earthquakes tend to cluster formation [6] around the epicenter of the main shock, thus these activities can be considered as a seismic that occurs at structural nodes. Seismicity patterns for earthquake prediction performance in appearance are useful to identify the physical mechanism [7]. When the physical mechanism is known, other tools such as a source mechanism changes, spectrum and waveform seismic records can be used for predictive purposes. Because of the heterogeneous catalogs available, earthquake seismic methods based on approximate reasoning help to track the behavior of complex processes. In practice, simulations such as artificial neural networks and Monte Carlo method and especially Copula to analyze the patterns make the distribution of seismicity on the fault plane more imminent.

Disorders of earthquake are other characteristics of fault systems that make it difficult to study them. The problem is that these systems together are peaceful coexistence of order and disorder. It seems that more complex systems such as the earthquake in the border of order and chaos are where to live. In this study, the criteria used to determine anomalies before earthquakes is more stable than environmental factors and performance measurement of nonlinear systems in the same land. This criterion is based on Copula. That is why the concept of Copula along with finding ways to fit it on earthquake data is analyzed and compared to environmental factors such as stability was evaluated. The results show the superiority of this criterion compared to other criteria's such as interdependence and mutual information. Besides, in this research, cluster map is used. Clustering 2-based method to produce the map of Monte Carlo algorithm is used. Finally, a functional network of different subjects is obtained by X-ray (graphical) analysis meaningful metrics such as clustering coefficient and length of Route 3 between earthquakes are compared.

From the innovative view, the use of new criteria (Copula) to measure the correlation between earthquakes, the use of the designated areas, and the use of clustering map to extract some features like the spatial correlation are noted for comparison. The suggested criteria for measuring correlation with other common measures in terms of performance have been compared. As well as the definition of high-risk areas of the clustering map, the catalog of seismic data obtained is used. The map does not have some restrictions in the use of probabilistic maps. Moreover, after the formation of the graphic simulation data, in addition to conventional metrics such as clustering coefficient and along the way, the metric used to compare the correlation of clustering coefficient has not been used for earthquake prediction.

1.1. Introduction of Copula theory

Copula functions of multivariate distributions to univariate marginal distribution functions are linked Copula literally. Copula could also be that borders univariate probability distribution functions are uniform. To clarify these definitions, consider the bivariate distribution: if X and Y are two random variables with marginal distributions $F(x) = P[X \le x]$ and $G(y) = P[Y \le y]$ the two combined cumulative distribution $H(x, y) = P[X \le x, Y \le y]$, then it can be any pair of real numbers (x, y). Three $G(y), F(x)$ and $H(x, y)$ are assigned to each of them between zero and one [0,1]. In other words, each pair (x, y) of real numbers with dots $(F(x), G(y))$ in the unit square $[0,1] \times [0,1]$ as well as the real number H (x, y) at domain [0,1] are corresponding.

Copula is the definition of a precise definition that has been mentioned in several studies. Figure (1) illustrates this definition better. High definition can be easily extended for distribution and multivariate. The important issues that will lead to a better understanding Copula scholar case, is described in the following section [8].

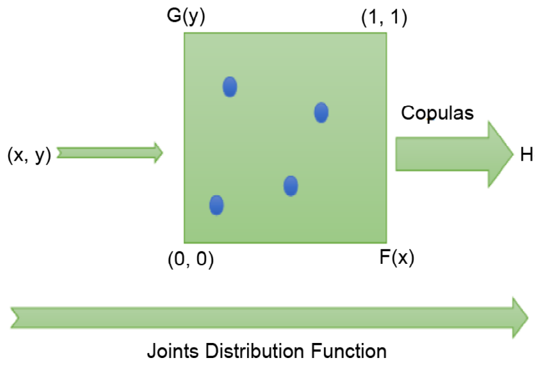

Figure 1. Copulas as relationship between outputs of combined cumulative distribution and marginal distributions.

1.2. Sklar's Theorem

If F distribution with margins F_1, F_2, ..., F_n (which are not necessarily continuous), then there is a Copula of n-dimensional space to a range of zero and one, $C:[0,1]n \to [0,1]$ that for all values of x_1, $x_2, \dots, x_n$ in the set of real numbers meet the following relationship:

$$F(x_1, \dots, x_n) = C(F(x_1), \dots, F(x_n)) \tag{1}$$

In this relationship, the marginal distribution functions are continuous, then C is unique; otherwise C on $Ran\,F_1 \times Ran\,F_2 \dots \times Ran\,F_n$ where Ran, shows a range of peripheral functions uniquely determined. It can be easily deduced from the following relationship:

$$C(u_1, u_2, \dots u_n) = \\ F((F_1^{-1}(u_1), F_2^{-1}(u_2), \dots, F_n^{-1}(u_n)) \tag{2}$$

This equation shows that C is a distribution function. Therefore, Copula multivariate distribution and its marginal distributions are uniform. According to the above description of any distribution function C, $C:[0,1]^n \to [0,1]$ with the following properties is a Copula:

$C(x_1, x_2, \dots, x_n)$ for each component of x_i is increasable.

For all amounts of i and with this condition:

$i \in \{1, \dots, n\}, x_i \in [0,1]$ then $C(1, \dots 1, x_i, 1, \dots, 1) = x_i$

For all amounts of a_i and b_i and with this condition:

$(a_1, \dots a_n), (b_1, \dots, b_n) \in [0,1]^n$, $a_i \le b_i$ then

$$\sum_{i_1=1}^{2} \dots \sum_{i_n=1}^{2} (-1)^{i_1+\dots+i_n} C(x_{1i_1}, \dots, x_{ni_n}) \ge 0$$

In this relationship for all amounts of $j \in \{1, \dots, n\}$ then $x_{j1} = a_j, x_{j2} = b_j$. Because the variables x_i actually obtained values of distributions is possibly proving up relations with this view will be very intuitive and simple. Copula's definitions show that statistical distributions as a parametric form such as normal distribution or parametric forms have been obtained experimentally. Unlike statistical distributions, finding suitable Copula, desired data can be directly acquired data dependency structure studied. Then, one of the most important families of parametric Copula is explained.

2. Archimedean Copula

In this section, a series of Copula named Archimedean is examined. Although Copula presented different parametric families, but this category of Copula is for very useful properties [3].

- Simplicity of making this category of Copula (simply fit the family's data).
- Many families fall into this category are parametric Copula.
- Diversity in the extraction and analysis of different attachment structures.
- Desirable characteristics of each member of the family of Copula.

Archimedean Copula makes it possible to study multivariate Copula in the form of a function to provide us supposed set of functions φ, where $\varphi:[0,1] \to [0,\infty]$ and φ is strictly decreasing continuous function, convex and values at zero and infinity, respectively. If φ^{-1} reverse of φ has the same properties, any member of φ Copula will produce Archimedean C function bivariate with margins uniform on $[0,1]$ derived from the following equation:

$$C(u,v) = \varphi^{-1}(\varphi(u)+\varphi(v)) \tag{3}$$

In this relationship, φ makes C. Simple example is multiplication function that named Archimedean Copula [3]. If X and Y are both independent, then for all amount of X and Y in the real Integer set:

Now if the generator of $\varphi(t) = -\ln(t)$ will be $t \in [0,1]$ then:

Therefore, multiplication function is Archimedean Copula [3]. To obtain the actual structure of dependency between statistical variables in various fields is very significant. Since the earthquake and seismic studies can be time-series statistical variables to be taken into account to determine the functional dependence between time series analysis, critical communications (location of future earthquakes) can be quite impressed from the results and analysis. If you have joint distribution of X_1 and X_2 variables, we can extract the dependence structure that lies in the distribution. In fact, with the help of conditional probabilities, marginal distributions can gain their behavior by combining distribution. However, along the distribution is often not a simple task. Copula extracts dependency structure without having to directly measure the distribution closer together for us. As stated by Nelsen, Copula is important for two main reasons: (1) as a method to study the structure demographic variables and (2) as a starting point in the construction of bivariate distributions (usually to generate random numbers). Copula and metrics are based on the favorable properties to determine the association between demographic variables including time series are: to understand the importance and use of standard-based performance measurement Copula affiliation, consider the limitations of cross-correlation criterion used in most studies. In theory, this standard only measures the linear dependence, whereas the time series of earthquake events can have a nonlinear relationship with each other. One of the disadvantages of mutual solidarity and many other metrics is that if variables X and Y are under strict increasing conversions, this measure may not stick, in fact, if ρ is the correlation between X, Y and T, a linear increase will be found: $\rho(T(X),T(Y)) \neq \rho(X,Y)$. This restriction is highly effective in comparison dependencies. For example, if X and Y variables are measured in gram or kilogram, obtained dependence between them in the case of X or Y or both measured on a logarithmic scale is different, because of applying the logarithm (that linear and incremental function) change to dependence. This criterion is based on Copula also higher accuracy in measuring nonlinear relationships shows. This ratio has good resistance to environmental factors. This is because the other criteria depend on measuring the correlation between each of their disadvantages. The properties are presented in this study if δ full measure to demonstrate the dependence of the variables X and Y, then you should have the following properties:

- This criterion should be symmetrical: $\delta\,(X, Y) = \delta\,(Y, X)$
- This criterion should be normalized: $1- \leq \delta(X, Y) \leq 1$.
- If each variable of X or Y are nonlinear function then $\delta\,(Y, X) = 1$.
- If α and β are ascending function on the variables of X and Y, then: $\delta(\alpha(X),\beta(Y)) = \delta(X, Y)$.
- $\delta(X, Y) = 0$ If only X and Y is independent.

Cross correlation only have first and second characteristics of them, but only Eq. (4) has been suggested by Schweitzer [3]:

$$\delta(X,Y) = 12 \times \iint_{I^2} |C(u,v) - uv|\,dudv \tag{4}$$

In this regard, C Copula fitted to the variables X and Y, as well as u, v as marginal distributions X, Y are shown respectively. In addition to the above equation based on other criteria defined by Copula, but this criterion was used in this study. The comparison of this method and other methods, such as interdependence and mutual information is indicative of more stability than noise. The results suggest that this method performs much better when the detection of communication is non-linear.

3. Processing the Locations of Future Earthquakes Based on Copula Simulation Algorithm

In view of the above, a natural approach to simulate from the Gauss Copula is to simulate from the multivariate standard normal distribution with an appropriate correlation matrix P, and to convert each margin using the probability integral transform with the standard normal distribution function. Whilst simulating from a multivariate normal distribution with covariance matrix Σ essentially comes down to do a weighted sum of independent standard normal random variables, where the "weight" matrix A can be obtained by the Cholesky decomposition of the covariance matrix Σ.

Therefore, an algorithm to simulate *n* samples from the Gauss Copula with correlation matrix *P* is:

- Perform a Cholesky decomposition of *P*, and set *A* as the resulting lower triangular matrix.
- Repeat the following steps n times.
- Generate a vector $Z=(Z_1,...,Z_d)'$ of independent standard normal variates.
- Set $X= AZ$
- Return $U=(\Phi(X_1),...,\Phi(Xd))'$.

Many areas of science depend on exploratory data analysis and visualization. The need to analyze large amounts of multivariate data raises the fundamental problem of dimensionality reduction: how to discover compact representations of high-dimensional data. Here, we introduce Locally Copula (LC), an unsupervised learning algorithm that computes low-dimensional, neighborhood-preserving embeddings of high-dimensional inputs earthquakes catalogue. Unlike clustering methods for local dimensionality reduction, LC maps its inputs into a single global coordinate system of lower dimensionality, and its optimizations do not involve local minima. For this reason, the concept of finding and fitting Copula refers to the earthquake data and earthquake, such as what was used clustering. The results show the superiority of this ratio has been on criteria such as artificial neural networks. In this paper, data pre-processing catalog of the aftershocks and earthquakes were used before. Then, using this algorithm of future seismic risk clusters (red area in Figures (2) and (3) and to find the aftershocks future (Star) is separated. In this graph, each node, which corresponds to the average area of these locations, indicates the distance between the nodes and reveals the correlation achieved between them from the proposed Copula-based criteria, Figure (2).

One of the great advantages of having statistical software like *R* available, even for a course in statistical theory, is the ability to simulate samples from various probability distributions and statistical models. This area is worth studying when learning *R* programming because simulations can be computationally intensive so that learning effective programming techniques is useful.

Using these posterior samples, we calculate cross-validation predictive densities to check our model. The 95% confidence intervals are displayed in Figure (3). One can see that only two out of 42 data points fail to be inside the corresponding

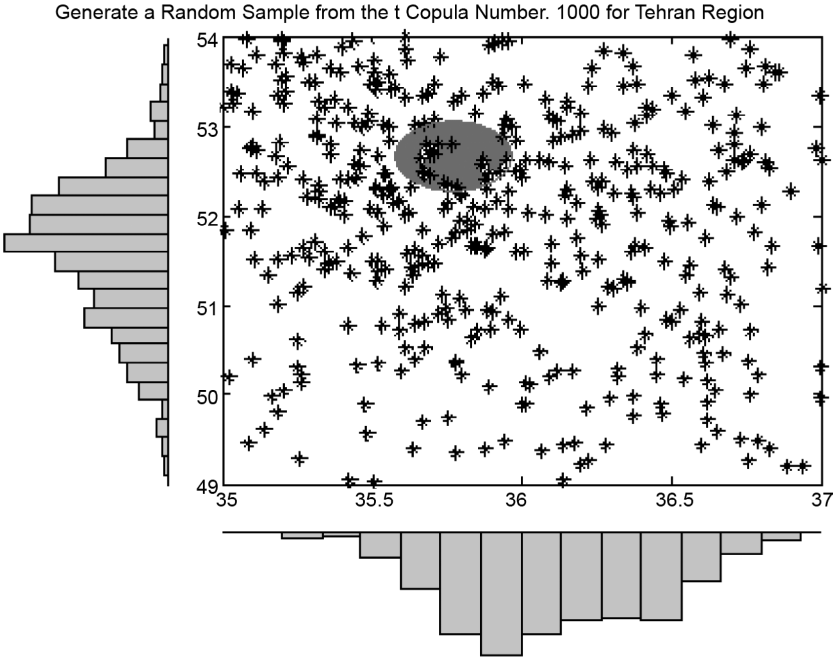

Figure 2. Doughnut pattern visualized by Copula's Network that is simulated for earthquake (2013/04/09, with Magnitude 6.3).

confidence interval. They are both located at the border of the sampling area and surrounded only by low values. The average of the standardized residuals is $r = -0.0012$.

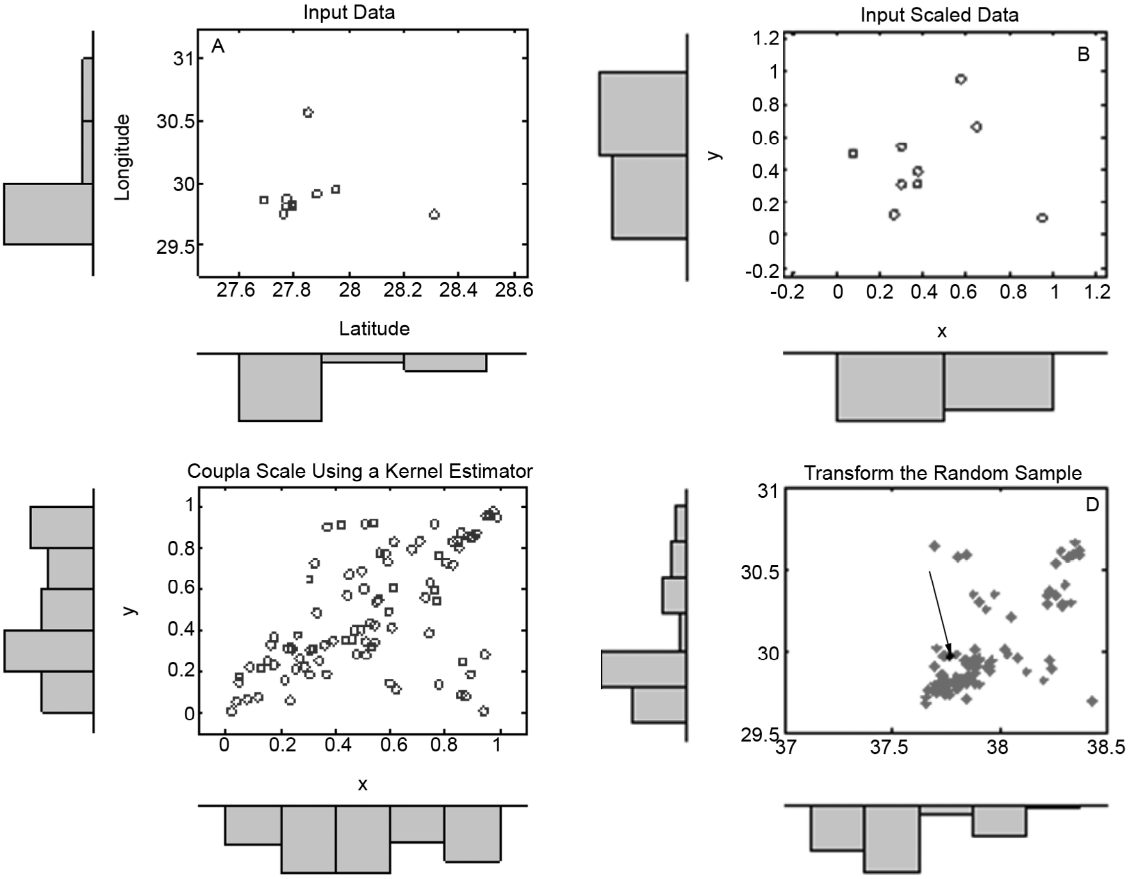

Figure 3. Doughnut pattern recognized by Copula's Algorithm for earthquake (1981/07/28, with Magnitude 7.3), doughnut pattern shows the place of event.

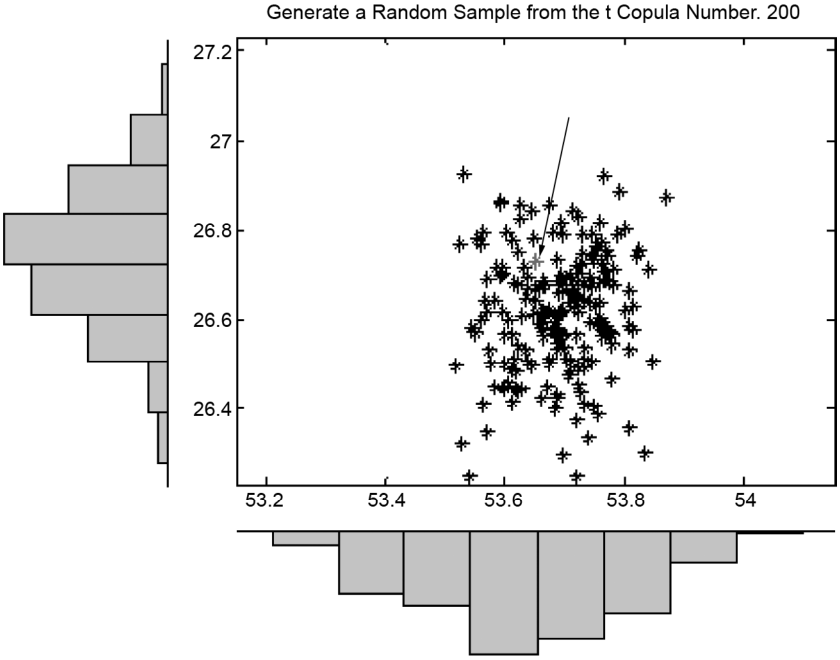

Figure 4. Doughnut pattern recognized by Copula's Algorithm for earthquake (1969/11/15, with Magnitude 7.3), doughnut pattern shows the place of event.

The mean squared error of cross-validation for the Copula methods and for the Self-organizing Neural Networks method is 16.34 and 18.55, respectively. Hence, both the Gaussian and the non-Gaussian model are not rejected by the data, but the Copula approach performs slightly better.

4. Conclusion

Earthquake forecasting based on extracting seismicity pattern can be a very difficult task. Various simulations methods and tools are used for detection of the precursor's seismic patterns such as doughnut pattern. In this work, we set Copula algorithms that have been shown to be powerful tools in various applications in seismology, and other areas. In this article, we use Copula densities to model class conditional distribution for pattern recognition of seismicity doughnut shapes before large earthquakes. These types of densities are useful when the marginal densities of event location pattern vector are not normally distributed. Those models are also useful for mixed pattern vectors. We also did simulations to compare the performance of the Copula-based classifier with classical normal distribution based model. Copulas offer an interesting opportunity to describe dependence structures for multivariate distributions. In recent years, new methods have been proposed for developing Copulas based on the self-organizing theory. These Copulas provide the opportunity to derive probability distribution function of multiple dependent variables and their dependant structure. The three categories of the Copula models are used in the type of numerical simulation such as Monte Carlo used for illustrating. The method uses the continuous functions of the marginal probabilities as the constraints to derive the Copula density function. The numerical example provided here, shows that the marginal probabilities and the dependence structure can be preserved reasonably well. One of the most important steps in the analysis of seismic data is to determine clustering and the correlation between them. So far, several criteria, such as artificial neural networks and SOM neural were used to determine these patterns, but the complexity and low efficiency of the environmental factors, the groundwork for further research to provide metrics that show the dependence structure provides a more precise overview. In this study, Copula-based criteria for identifying high-risk clusters and patterns presented doughnuts better performance than self-organizing neural networks has shown. There is a very simple method to simulate from the Gaussian Copula that is based on the definitions of the multivariate normal distribution. This indicates that, in a practical application, one can improve the model performance by increasing the number of discrete points. Due to the limitations of some of the criteria in determining the relationships of earthquakes, the research-based criteria used to determine relationships between different areas are used. It seems that the following criterion that was compared with some common standards in this area looks much better. In Figure (3), the steps achieved on Copula about the great earthquake of (7.3) July 28, 1981 (1360/05/06) in the region occurred in the Kerman province can be seen. Figure (3-A) is the input or foreshocks before. Figure (3-B) using a kernel estimator Copula scale applied then in Figure (3-C) using Monte Carlo simulation data increases. In the synthetic data with production Copula (10x10) pattern visualize doughnuts patterns before it is occurred. Earthquake groups joined together in this way, in fact, to quantify and identify groups that are interrelated earthquake probabilities are very high and in the future.

References

1. Field, E.H., Dawson, T.E., Felzer, K.R., Frankel, A.D., Gupta, V., Jordan, T.H., Parson, T., Petersen, M.D., Stein, R.S., Weldon II, R.J., and Wills, C.J. (2009) Uniform California eathquake rupture forecast, Version 2 (UCERF2). *Bulletin of the Seismological Society of America*, **99**(4), 2053-2107.

2. Allamehzadeh, M. and Mostafazadeh, M. (2014) Determination of concentration of earthquake clustering. *7th International Conference on Seismology of Earthquake Enginerign (SEE7).*

3. Nelsen, R.B. (2006) *An Introduction to Copulas.* Springer Series in Statistics.

4. Joe, H. (1997) Multivariate models and dependence concepts. Vol. 73 of Monographs on Statistics and Applied Probability, Chapman & Hall, London, UK.

5. Jones, L.M. and Molnar, P. (1976) Frequency of foreshocks. *Nature*, **62**, 677-679.

6. Mogi, K. (1968) Development of aftershock areas of great earthquakes. *Bull. Earthq. Res. Inst.*, **46**, 175-203.

7. Allamehzadeh, M., Mostafazadeh, M., and Mahshadnia, L. (2013) *Developed Sophisticated Pattern Recognition of Earthquake Location, Simulation Alborz Region*. Report Project No. 9604-93-6 at IIEES.

8. Pitt, M., Chan, D., and Kohn, R. (2006) Efficient bayesian inference for Gaussian Copula regression. *Biometrika*, **93**, 537-554.