JSDS: JOURNAL OF STATISTICS AND DATA SCIENCE
VOLUME 02, No 02, October 2023
e-ISSN: 2828-9986
https://ejournal.unib.ac.id/index.php/jsds/index

Earthquake Clustering Using the CLARA Method and Modeling Using the Inhomogeneous Spatial Cox Processes Method in the Ambon Region

Muhamad Iqbal Meiwidian [1], Riki Crisdianto [1*], Dyah Setyo Rini [1]
[1] S1 Statistics Study Program, Bengkulu University, Bengkulu

* Corresponding Author: rikic628@gmail.com

Article Info

Article History:
Submited: October 1, 2023
Accepted: October 23, 2023
Available Online: October 30, 2023

Key Words:
Earthquake
Ambon
CLARA
Neyman-Scott Cox Process
K-function

Abstract

Earthquakes are natural events whose time and place cannot be predicted. Ambon is the largest city in the Maluku Islands region and is the center of development and the capital of Maluku Province. This research will group earthquake events, analyze the characteristics of earthquake events, create earthquake zones and map them using CLARA cluster analysis, and create modeling that will look at the risk of earthquake events in a location based on distance to faults and subduction zones using the Inhomogeneous Neyman-Scott Cox Process. The data used is data on earthquake events in the Ambon region obtained from the United States Geological Survey (USGS) catalog from January 1926 to December 2022, with a depth of ≤360.1 Km and a magnitude of ≥4 Mw. Grouping earthquake events in the Ambon area using CLARA cluster analysis obtained 2 groups of earthquake clusters with an optimal silhouette score of 0.7430. The model obtained in this earthquake research is not good because it is based on the K-function value plot of the original data which is far from the modeling K-function value plot.

1. INTRODUCTION

Earthquakes are genuine vibrations from within the earth, originating within the earth which then propagate to the earth's surface due to cracks in the earth breaking and shifting violently [1]. Meanwhile, according to Budianto et al [2] earthquakes are one of the natural disasters that often occur in Indonesia, because Indonesia is located in the world's ring of fire. The time and place of an earthquake cannot be predicted. Although earthquakes are prone to occur in the ring of fire, the location and time when they will occur cannot be scientifically determined. In general, most of the Indonesian archipelago is located in the ring of fire area which stretches from the islands of Sumatra, Java, Bali, Nusa Tenggara, Sulawesi, Maluku, to Papua. Therefore, areas in the territory of the Unitary State of the Republic of Indonesia can be stated to be prone to earthquake hazards [3]. On the other hand, Indonesia is also an archipelagic country located between three tectonic plates, namely the Australian, Eurasian and Pacific plates. This condition places Indonesia as one of the countries that is vulnerable to natural disasters.

Ambon is the largest city in the Maluku Islands region and is the center of development and the capital of Maluku Province. Ambon has a fairly high potential threat of earthquakes and tsunamis. Although information is limited, there are written historical records that provide some information. The first fairly complete record of the earthquake and tsunami in Indonesia was the earthquake and tsunami on Ambon Island which occurred on February 17, 1674 [4]. This earthquake and tsunami is estimated to have claimed up to 2,500 lives. Due to the earthquake and tsunami disaster conditions, Ambon City is very at risk of the impact of a tsunami.

From a spatial statistics point of view, there is a relationship between earthquake occurrences between points of occurrence, where the results are statistical there is a close relationship between earthquake occurrences between location points. Statistically, the proximity and characteristics of these earthquake occurrence points can be grouped using a cluster analysis approach. Cluster analysis is a multivariate grouping method with the main aim of grouping subjects or objects based on their similar characteristics. Cluster analysis has high homogeneity (similarity) between members in one group, and high heterogeneity (difference) between one group and another group.

This research will group earthquake events, analyze the characteristics of earthquake events, create earthquake zones and map them using CLARA cluster analysis, and create modeling that will look at the risk of earthquake events in a location based on distance to faults and subduction zones using the Inhomogeneous Neyman-Scott Cox Process.

1.1 Clustering Large Application (CLARA)

Clustering large application (CLARA) is an algorithm that can work effectively for large datasets, this algorithm is sampling-based. Thus, the partition method can still be carried out based on the principle of minimizing the number of dissimilarities between each object and the corresponding reference points (medoids).

In contrast to the PAM medoid method, the CLARA method can be more efficient in terms of computing time and in storing large datasets. CLARA applies the PAM algorithm to obtain the optimal medoid for the sample. The quality of the resulting medoids is measured by the average difference in distance between each object in the dataset and the medoids in the sample. By taking samples randomly, the medoid of the sample is expected to be close to the medoid value of the dataset. The CLARA algorithm is an algorithm that overcomes the weakness of the k-means algorithm which is sensitive to outliers because objects with a large value may deviate from the data distribution.

The steps of the CLARA clustering method include [5]:

1. Initialize k cluster centers (number of groups).
2. Allocate each data (object) to the closest cluster using the Euclidian Distance distance measure equation with the equation:

$$d_{ij} = \sqrt{\sum_{a=1}^{p} \left(x_{ia} - x_{ja}\right)^2} = \sqrt{\left(x_i - x_j\right)'\left(x_i - x_j\right)} \tag{1}$$

Where $i = 1, \ldots\ldots, n;\ j = 1, \ldots\ldots, n$ and p are the number of variables.

3. Randomly select an object in each cluster as a new medoid candidate.
4. Calculate the distance of each object to the new medoid candidate and assign each object to the nearest medoid candidate.
5. Calculate the total distance, then the difference in the total distance "(S)" by calculating the new total distance value - the old total distance. If "S < 0", then swap the objects with the cluster data to form a new set of k objects as medoids.
6. Repeat steps 3 to 5 until no medoid changes occur, so you get clusters and their respective cluster members.

1.2 Distance Measurements

Euclidean distance measurements will be carried out for the CLARA and k-means algorithms. Euclidean distance has several advantages, including the distance between two objects is not affected by the addition of new objects to be analyzed, which may be outliers [6]. In euclidean distance, the distance between two points is defined as a straight line. Euclidean distance can be calculated by

$$D_{Euclidean}\left(x_i, x_j\right) = \sqrt{\left(x_i - x_j\right)^2} = \sqrt{\sum_{m=1}^{n} \left(x_{im} - x_{jm}\right)^2} \tag{2}$$

$D_{Euclidean}\left(x_i, x_j\right)$: Euclidean distance
x_i : The i-th data
x_j : The j-th data
x_{im} : The i-th data of the m-th attribute
x_{jm} : The j-th data of the m-th attribute

In research conducted by Mohibullah [7] who tested the Euclidean and Manhattan distances using CLARA, it was concluded that calculating the two distances gave equally good results but still depended on the input data. If you use the Euclidean method, it is better to use small secret data, while if you use the Manhattan method, it is better to use large secret data.

1.3 Silhoutte Score

In clustering, determining the number of clusters is one of the most important steps. Silhouette coefficient is a method used to evaluate clusters and see the quality of data placement in a cluster. The silhouette score is important to see whether the resulting cluster is of good quality. The stages for calculating the Silhoutte coefficient are as follows [8]:

1. Calculate the average distance of object i to all objects in the group. We call the average distance $a(i)$.
2. Calculate the average distance of object i to all objects in the other cluster which we call $b(i)$, and take the smallest value.
3. The Silhoutte coefficient value is obtained by

$$S(i) = \frac{b(i) - a(i)}{max\ (b(i), a(i))} \tag{3}$$

Can be written

$$S(i) = \begin{cases} 1 - \frac{a(i)}{b(i)} & , jika\ a(i) < b(i) \\ 0 & , jika\ a(i) = b(i) \\ \frac{b(i)}{a(i)} - 1 & , jika\ a(i) > b(i) \end{cases}$$

Where:

$S(i)$: silhouette score

$a(i)$: the average distance between data i and all objects in the cluster.

$b(i)$: the average distance of data i to all objects in other clusters

The range of values for the silhouette coefficient is -1 to 1. If the value of the silhouette coefficient is close to 1 then the object is in the right cluster, if it is around 0 then the object could be between 2 clusters, and if the result is negative then the object may be in the wrong cluster. The best number of clusters or optimal number of clusters is the number of clusters with the highest average silhouette score where the average is taken from the silhouette value of each cluster.

1.4 Inhomogeneous Neyman-Scott Cox Process

We can model the risk of an earthquake that will occur using the Inhomogeneous Neyman-Scott Cox Process, which is a form of the Cox Process, where C is the main earthquake generated from a stationary Poisson process with intensity $k > 0$. Then $X_{c,c\epsilon C}$ is an aftershock from an independent Poisson process with the following intensity function.

$$\rho_c(u;\beta) = \exp\ (\beta^T x(u))k(u - c;\omega)$$

Where, k is the probability density function of the distance distribution between aftershock and main earthquake with parameter ω, then $X = U_{c\epsilon C} X_c$is an Inhomogeneous Neyman-Scott Cox Process with intensity function $\rho(u;\beta)$ in the following equation.

$$\rho(u;\beta) = k\ \exp\ (\beta^T x(u))$$

So the formula for the Pair correlation function Thomas Cluster Process is obtained in the following equation.

$$g(u) = 1 + (4\pi\omega^2)^{-1} \exp\left(-\frac{\|u\|^2}{4\omega^2}\right)/k \tag{4}$$

When the value of ω gets smaller, the distance between aftershocks and main earthquakes gets shorter, so the clusters formed tend to be tight, and if the value of k gets smaller, then this results in the number of main earthquakes in a region will be less because it is proportional to the smaller value of k [9].

1.5 K-function

The goodness of the model is seen based on the K-function envelope plot. The original concept of K-function is to calculate the distance between all pairs of different points and the cumulative average-homogeneous sum derived from the distance between points that is less than or equal to the value of r. Spatial relationships are used to select a structured division of spatial point patterns whether regular or points tend to avoid each other, independent, and clustering (between points tend to be close together or in groups) [10]. A model is said to be good for modeling data if the original data K-function value plot is in the K-function envelope interval [11].

1.6 Earthquake

An earthquake is an event of shaking or shaking of the earth due to sudden movement/shifting of rock layers on the earth's skin due to the movement of tectonic plates [12]. Earthquakes can occur as a result of faulting. The kawa horizontal fault structure itself has an arc-shaped pattern, there has been a series of small earthquakes that occurred around the fault, and this is the opening earthquake before the big earthquake occurred. Therefore, the presence of this active fault is important to be identified for future seismic hazard assessment. The kawa horizontal fault structure itself has an arc-shaped pattern, curving north from east to west, stretching a distance of up to 453 km. Before the 6.5 magnitude earthquake occurred, BMKG noted, since August 28, 2019, there had been a series of small earthquakes that occurred around the fault, and this was the opening earthquake before the big earthquake occurred. Ambon, Haruku, and Kairatu earthquakes.

1.7 Ambon Island

Ambon Island is characterized by a geomorphology dominated by structural mountains. Ambon is located at "3034'4.80" - 3047'38.4" LS" and 12801'33.6"-128018'7.20" East and there is a border with the Banda Sea causing the city to be located in the mid-Asiatic region which is part of the Eurasian Plate. In the last decade, the 2019 Ambon Earthquake occurred on September 26, 2019 at 06:46 WIB with a magnitude of 6.6 Mw(mB) according to the Meteorology, Climatology and Geophysics Agency [13]. The epicenter of the earthquake was located at coordinates-3.42°N,128.45°BT at a depth of 10 km according to the BMKG update. This earthquake was felt by the community with a maximum intensity of VI MMI (Modified Mercally Intensity) and caused damage to infrastructure and community homes, and claimed the lives of at least 30 people. The earthquake could have been generated by an east-west trending structure crossing the southern Kairatu area that generally corresponds to strike slip lineations in the northern and southern parts of Seram Island, or it could have been generated by a north-south trending structure that has not yet been identified. The strike-slip and thrust type faulting around Seram Island is tectonically natural as a result of intense deformation associated with the convergence of the Australian, Eurasian and Pacific plates and several microplates with different scales and velocities. Several strike slip and thrust structures in this region have been identified previously and have been included in the 2017 Indonesia Earthquake Hazard and Source Map [14].

2. METHOD

The data used in this study are earthquake events in the Ambon region (January 1, 1926 to December 31, 2022) obtained from the United States Geological Survey (USGS) website. In detail, the stages of this research are as follows:

1. Literature Review and Methodological Exploration
2. Data collection
3. Creating data visualization to determine the characteristics of earthquake distribution in the Ambon region.
4. Performing CLARA cluster analysis
5. Creating data visualizations from the results of CLARA cluster analysis to determine the characteristics of earthquake distribution in the Ambon region.
6. Modeling earthquake data on Ambon island with inhomogeneous Neyman-Scott Cox Process.
7. Checking the goodness of the model by looking at the K-function envelope plot.

3. RESULTS AND DISCUSSION

3.1 Earthquake Characteristics

During 1926-2022, 807 earthquakes occurred in the Ambon island region. The following is the distribution of longitude 127 to 129.5 and latitude -4.3 to -2.8 earthquakes that occurred near the Ambon area.

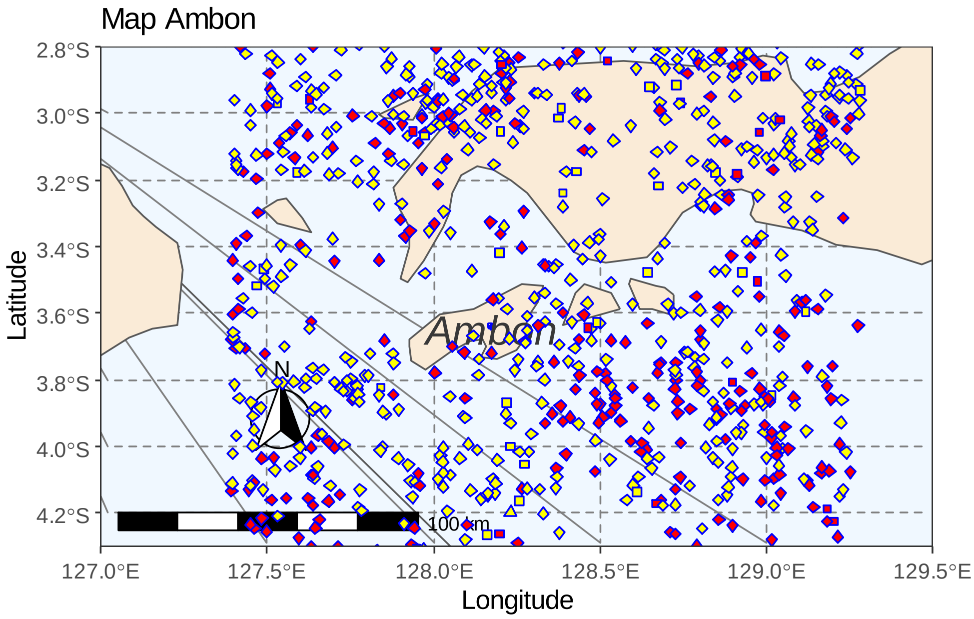

Figure 1. Distribution of Major Earthquakes on Ambon Island 1926-2022

Based on Figure 1, it can be seen that most of the earthquake events in the Ambon region occurred in marine waters. Earthquake events with depths of less than 50 km (<50 km) are symbolized by yellow dots, while red dots are earthquake events with other depths of less than 300 km. In other words, shallow and deep earthquakes are spread over both land and sea waters. Descriptive statistics of earthquake occurrences in the Ambon region can also be seen in the following table:

Table 1. Descriptive Statistics of Earthquake Depth and Magnitude in the Ambon Region

Descriptive Statistics	Depth	Magnitude
Average	57.24	4.563
Median	35	4.5
Mode	33	4.2
Standard Deviation	51.97124	0.5518751
Minimum	4	3
Maximum	360.1	7.3

Based on table 1, earthquake events in the Ambon region have an average depth of 57.24 and an average magnitude of 4.563, have a median depth of 35 and a median magnitude of 4.5, have a mode depth of 33 and a mode magnitude of 4.2, have a standard deviation of depth of 51.97124 and a standard deviation of magnitude of 0.5518751, have a minimum depth of 4 and a minimum magnitude of 3, have a maximum depth of 360.1 and a maximum magnitude of 7.3.

3.2 CLARA Cluster

Grouping earthquake events in the Ambon region using CLARA cluster analysis obtained 2 earthquake cluster groups with an optimal silhouette score obtained of 0.7430. CLARA clusters in grouping earthquake data on the island of Ambon can be seen from Figure 2.

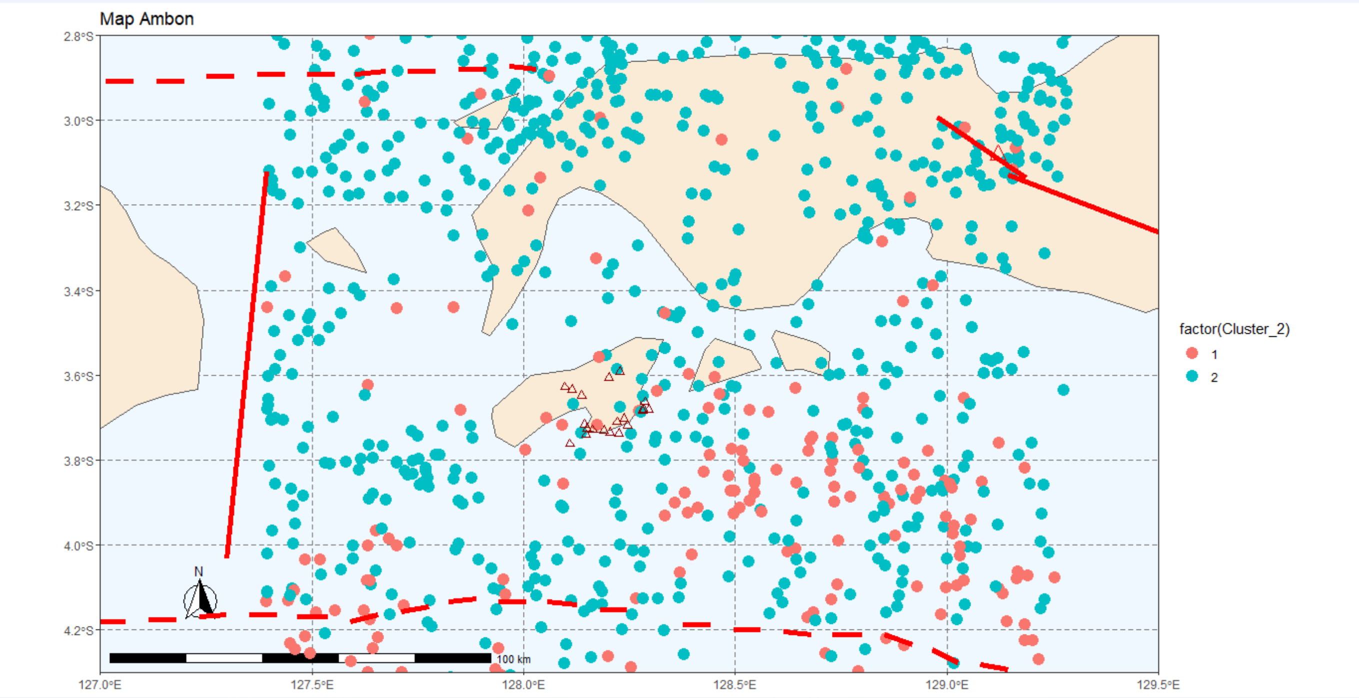

Figure 2. Clusters of Large Earthquakes on Ambon Island 1926-2022

Here are some characteristics of the resulting clusters:

Table 2. Characteristics of Ambon Region Earthquake Clustering Results

Cluster	Average Latitude	Average Longitude	Average Depth	Average Magnitude	Number of Members
1	-3.8595	128.445	143.474	4.48661	165
2	-3.3505	127.954	35.0728	4.58224	642

Cluster 1 has an average latitude of -3.8595 and an average longitude of 128.445. Cluster 1 has 165 members, an average magnitude of 4.48661 Mw, and an average depth of 143.474 Km. The earthquakes that occur in this cluster are classified as medium earthquakes, as can be seen from table 2. This cluster has fewer earthquake events than cluster 2 earthquake events, it can be seen from the small number of earthquake events.

Cluster 2 has an average latitude of -3.3505 and an average longitude of 127.954. This cluster has an average magnitude of 4.58224 Mw, an average depth of 35.0728 Km, and a total membership of 642 earthquake points. The earthquakes that occur in this cluster are classified as shallow earthquakes, as can be seen from table 2. This cluster has many earthquake events, this can be seen from the number of earthquake events that are so dominant when viewed in table 2.

3.3 Chi-Square Test

The Chi-Squared test is conducted to determine whether the earthquake data follows a stationary pattern or not. Based on the output of the chi squared stationarity test in the appendix, with testing using a significance level of 0.05, a p-value of $< 2.2e - 16$ was obtained. Because the p-value <0.05, it is concluded that the earthquake data on the island of Ambon follows a non-stationary pattern, meaning that there may be factors that significantly affect the number of earthquakes that cause the number of earthquake events in each region of Ambon to be different.

3.4 Earthquake Data Modeling

The results of modeling earthquake data on the island of Ambon using the inhomogeneous Neyman-Scott Cox Process model with a significance level of 5% concluded that the geographical factor of fault and volcano distance significantly affects the risk of earthquake occurrence on the island of Ambon. This means that the risk of earthquakes in each region in Ambon is different, depending on the distance of the area to faults and volcanoes. The estimator values for the parameters κ, β and ω are presented in Table 3.

Table 3. Estimated Values of $\hat{\kappa}$, $\hat{\beta}$ and $\hat{\omega}$

Coefficient	Estimate	1/EXP (Beta)
$\hat{\kappa}$	5.161561e-10	
$\hat{\omega}$	5.810521e-01	
$\hat{\beta}_0$	3.9299267	0.01964511
$\hat{\beta}_1$	0.8546314	0.42544
$\hat{\beta}_2$	-1.4742066	4.367569

Table 3 shows that the value of $\hat{\kappa}$ obtained from the estimation results is 5.161561e^{-10} and $\hat{\omega}$ is 5.810521e^{-01}. After estimating the parameters, the intensity model of the inhomogeneous Neyman-Scott Cox Process is obtained as follows:

$$\hat{\rho}(u) = 5.161561\text{e}^{-10} \times \exp\,(3.9299267 + 0.8546314\text{X}_1(\text{u}) - 1.4742066\text{X}_2(\text{u})) \quad (5)$$

From the estimated model in equation 5, information can be obtained that:

1. If the distance between a location and faults and volcanoes is 0, then the probability of an earthquake occurring at that location is 0.01964511.
2. If the distance from a location to a volcano is fixed, but the distance from the location to a fault increases by one unit, then the probability of an earthquake occurring at that location is 0.42544 times greater.
3. If the distance from a location to a fault is fixed, but the distance from the location to a volcano increases by one unit, then the probability of an earthquake occurring at that location is 4.367569 times smaller.

3.5 Model Goodness Check

The goodness of the inhomogeneous Neyman-Scott Cox Process model in modeling earthquake data on the island of Ambon can be seen from the K-function envelope plot in Figure 3.

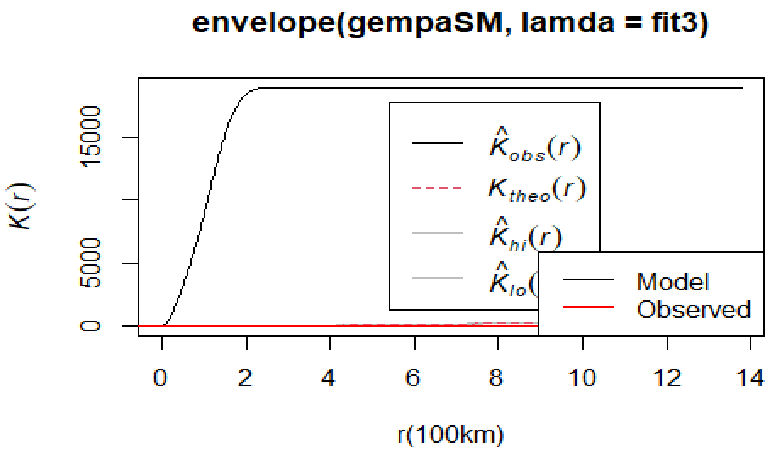

Figure 3. Envelope Plot of Inhomogeneous Neyman-Scott Cox Process Model

Figure 3 shows that the K-function plot of the earthquake data is far from the K-function envelope interval. This can be seen from Figure 3, where the straight black line is in the upper area, meaning that the inhomogeneous Neyman-Scott Cox Process is not good enough to be used to model earthquake data in the Ambon region in 1926-2022 with longitudes 127 to 129.5 and latitudes -4.3 to -2.8. This can happen because the average earthquake event in the Ambon region has similar characteristics. It is recommended that future research use other methods, such as RSA (Response Spectral Analysis) which is useful when modeling earthquakes with similar characteristics in certain structures [15].

4. CONCLUSION

The Ambon region, located in eastern Indonesia, is one of the areas that frequently experiences earthquakes. The characteristics of earthquake events in the Ambon region mostly occur in marine waters. Earthquakes with depths of less than 50 km (<50 km) are symbolized by yellow dots, while red dots represent earthquakes with other depths of less than 300 km. In other words, shallow and medium-sized earthquakes are spread over both land and sea waters. Grouping earthquake events in the Ambon region using CLARA cluster analysis obtained 2 earthquake cluster groups with an optimal silhouette score obtained of 0.7430.

The chi-square test uses a significance level of 0.05, with a p-value of $< 2.2e-16$. Because the $p-value < 0.05$, it is concluded that the earthquake data on the island of Ambon follows a non-stationary pattern, meaning that there may be factors that significantly affect the number of earthquakes that cause the number of earthquake events in each region of the island of Ambon to differ.

Based on the analysis results, the intensity model of the inhomogeneous Neyman-Scott Cox Process is obtained as follows:

$$\hat{\rho}(u) = 5.161561\mathrm{e}^{-10} \times \exp\,(3.9299267 + 0.8546314\mathrm{X}_1(\mathrm{u}) - 1.4742066\mathrm{X}_2(\mathrm{u}))$$

Based on Figure 3, the K-function plot of the earthquake data is far from the K-function envelope interval. This can be seen from Figure 3, where the straight black line is in the upper area, meaning that the inhomogeneous Neyman-Scott Cox Process is not good enough to be used to model earthquake data in the Ambon region in 1926-2022 with longitudes 127 to 129.5 and latitudes -4.3 to -2.8.

ACKNOWLEDGMENTS

Thank you to the supervisors who have guided this research and to the S1 Statistics Study Program FMIPA Bengkulu University for funding this research..

REFERENCES

[1] A. M. Nur, “Gempa Bumi Tsunami dan Mitigasinya”, Jurnal Geografi, vol. 7, no. 1, pp. 66-73, Januari 2010.

[2] M. H. Budianto, I. W. Yasa, E. Setiawan, D. widianty, and R. Yuniarti, “Sosialisasi Tentang Bersahabat Dengan Bencana Alam Gempa Bumi Di Desa Pemenang Barat”, Jurnal Bakti Nusa, vol. 2, no. 1, pp. 15-22, April 2021.

[3] I.W. Subagia, I. G. L. Wiratma, and I. K. Sudita, “Pelatihan Mitigasi Bencana Alam Gempa Bumi Pada Siswa Sekolah Dasar Negeri 1 Pengastulan Kecamatan Bulelang Bali”, Jurnal Pendidikan Indonesia, vol. 4, no. 1, April 2015.

[4] I. Kurniawati and R. M. Rahman, “Teknik Pembuatan Peta Evakuasi Tsunami Negeri Galala & Hative Kecil, Kota Ambon”, Jurnal Unhas, vol. 8, no. 3, pp. 225-231, Juni 2021.

[5] A.P. Fialine, D.A. Alodia, D. Endriani, and E. Widodo, “Implementasi Metode K-Medoids Clustering untuk Pengelompokan Provinsi di Indonesia Berdasarkan Indikator Pendidikan”, Journal of Mathematics Education and Applied, Vol. 02, No. 02, pp. 1-13, November 2021.

[6] Novianti, P. Dyah, S. and Rafflesia, U., "K-Means cluster analysis in earthquake epicenter clustering", International Journal of Advances in Intelligent Informatics vol. 3, no. 2, pp. 81-89, 2017.

[7] M. Mohibullah, M. Z. Hossain, and M. Hasan, "Comparison of Euclidean distance function and manhattan distance function using k-mediods," International Journal of Computer Science and Information Security, vol. 13, no. 10, pp. 61-71, 2015.

[8] D. F. Pramesti, M. T. Furqon, and C. Dewi, “Implementasi Metode K-Medoids Clustering Untuk Pengelompokan Data Potensi Kebakaran Hutan / Lahan Berdasarkan Persebaran Titik Panas (Hotspot),” Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, vol. 1, no. 9, Juni 2017.

[9] R. Metrikasari and A. Choiruddin, “Pemodelan Risiko Gempa Bumi di Pulau Sumatera Menggunakan Model Inhomogeneous Neyman-Scott Cox Process”. JURNAL SAINS DAN SENI ITS Vol. 9, No. 2, pp. D102-D107, 2020.

[10] A. Baddeley, E. Rubak, and R. Turner, Spatial Point Patterns Methodology and Applications with R, 1st ed. London: Chapman and Hall/CRC, 2015.

[11] A. Baddeley, E. Rubak, and R. Turner, Spatial Point Patterns: Methodology and Applications with R, 1st ed. Boca Raton: CRC Press, 2016.

[12] S. Atmojo and I. Muhandis,“Sistem Informasi Geografis Bencana Gempa Bumi Dengan Pendekatan Pga Untuk Mitigasi Bencana”, Jurnal Ilmiah Edutic, vol. 6, no. 1, pp. 2407-4489, November 2019.

[13] G. Ibrahim and Subardjo, Buku Gempabumi, Penerbit BMKG, 2001.

[14] Sianipar, Dimas, Halauwet, Yehezkiel, dan Daryono, “Studi Sumber Gempa Ambon 26 September 2019 Mw 6,5”, 2019.

[15] Clough, R.W., and Penzien, J. (1993). "Dynamics of Structures".