EPJ Web of Conferences **68**, 00005 (2014)
DOI: 10.1051/epjconf/20146800005

Spatial Data Clustering and Zonation of Earthquake Building Damage Hazard Area

Irwansyah. E[1] , and Winarko.E[2]

[1]Dept of Computer Science, Bina Nusantara University, K.H. Syahdan No.9, Palmerah, Jakarta 11480 Indonesia
[2]Dept of Computer Science and Electronic, Universitas Gadjah Mada, Sekip Utara, Bulak Sumur , Yogyakarta 55281, Indonesia

Abstract. Research conducted with the aim to develop spatial data cluster and analyse the characteristics of each cluster of data in order to develop spatial zonation of building damage hazard caused by the earthquake in the city of Banda Aceh. The data used consists of the data peak ground acceleration (PGA), lithology and topographic zone data with the four phases of the study consisted of data normalization with min-max method, determination of optimum data group with the connectivity index, Dunn index and the Silhouette index, clustering of data with the *k*-Means algorithm and the data interpolation using kriging algorithm to construct the zonation of hazard area. The optimum numbers of clusters used in these researches are two, with the first 170 data largely concentrated at a distance of less than 0.1 to 0.5 from the cluster centre and the second 55 others data concentrated at a distance of 0.3 - 0.5 from the cluster centre. Banda Aceh and surrounding areas can be spatially divided into two classes of potential building damage caused by the earthquake with high hazard class building damage are generally located near the coastline.

1 Introduction

Since 1900, the United State Geological Survey (USGS) have noted that four major quake in Indonesia, Banda earthquake (Mw 8.5) in 1938, an earthquake of Sumatra-Andaman Islands (Mw 9.1) 2004, North Sumatra earthquake/ Nias (Mw 8.6) in 2005 (USGS, 2009) and the West Coast of Sumatra earthquake (8.6) in 2012 (USGS, 2012).

In addition to the large magnitude earthquakes (> 5MW), parts of Indonesia also happens almost every time an earthquake with a smaller magnitude. Indonesia meteorological, climatology and geophysics office (BMKG) notes that in Indonesia, almost every day there was an earthquake with a magnitude more than 5 MW (destruction earthquake) and every nine days there was an earthquake with the same magnitude in island of Sumatra region. Type of destruction earthquake, besides causing loss of life, also cause damage of the physical infrastructure, especially buildings. Total building damage reaches 35 percent as a result of the 2004 Sumatra earthquake (Irwansyah, 2010), more than 140,000 buildings are damage in Bantul Yogyakarta earthquake in 2006 (Miura et al, 2005) and around 300,000 collapse buildings as impact of the West Sumatra earthquake in 2009 (AusAID, 2012).

Impact assessment of building damage caused by the earthquake, at this time conducted by the researchers based on ground conditions using earthquake magnitude data and converted to the peak ground acceleration (PGA) value (Sengara, 2008; Irwansyah and Winarko, 2013), the lithology condition and topographic zone in addition based on data of building structures in the area are assessed (Tesfamariam and Saatcioglu, 2010).

Impact assessment process of the damage caused by the earthquake, the implementation involves large amounts of data, especially data in time series and also in higher dimensions. The processing of large amounts of data in high-dimensional and requires a proper way in order to produce a better conclusion. One common way is to perform data clustering based on similarity and dissimilarity relations between variables data. This way of grouping in data mining is known as cluster analysis.

Han et al, 2001, divides data clustering method into four major groups, namely (1) partitioning methods with the k-Means algorithm, expectation maximization (EM) and *k*-medoid, (2) hierarchical methods with the algorithm is known as AGNES, DIANA, CURE, BIRCH CHAMELON and, (3) density-based methods with the algorithm such as DBSCAN, OPTICS and DENCLUE and (4) grid-based methods with STING algorithm, WaveCluster and Clique. Partitioning algorithm *k*-Means algorithm is popular to this day because of its ease to implement and efficient for processing large amounts of data (Jain, 2009) with weakness in high sensitivity to noise and data is data that is outliner (Han et al, 2001) .

Steinbach et al, 2000 showed empirically that *k*-Means algorithm is good or better than the hierarchical algorithm.

Clustering results of the seismic data with only clustering method is not able to show the spatial patterns due to the randomly distribution of data and the low data density. Irwansyah and Hartati, 2012 implements kriging algorithm to predict the data are not available in some locations in order to construct spatial pattern of the earthquake hazard area in the city of Banda Aceh. Partitioning *k*-Means algorithm is used in this study to classify and analyse the pattern of damage data for each cluster and kriging algorithm to interpolate unavailable data in some location for preparation of the spatial patterns of each data cluster.

2 *k*-Means algorithm: development and application in spatial data

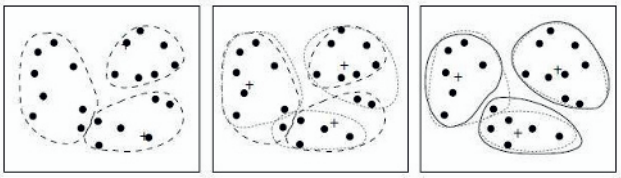
k-Means is a clustering algorithm with partitioning method is based on the central point (centroid) in contrast to *k*-Medoids algorithm is based objects. This algorithm was first proposed by MacQueen (1967) and developed by Hartigan and Wong 1975 with the goal of being able to share the M data points in N dimensions into a number of *k* clusters where the clustering process is done by minimizing the sum squares distance between the data with each cluster centre (centroid-based).

Partitioning *k*-Means algorithm in its implementation requires three parameters are entirely user-defined including number of clusters *k*, the initialization cluster, and the distance metric. *k*-Means usually run independently with different initialization produce different final clusters as the algorithm is in principle only classify the data towards local minima. One way to overcome the local minima is to implement the *k*-Means algorithm, for a given K, with some initial value of the different partitions and then selected the partition with least square error (Jain, 2009).

k-Means is a technique that is quite simple and fast in object clustering process. In the application of *k*-Means algorithm, when given a set of data X = {x1, x2, …,xn} in where xi = (xi1, xi2, …, xin) is a vector in real space Rn, then the *k*-Means algorithm will construct a partition X in the number of *k* clusters (a priori). Each cluster has a center point (centroid) which is the average value (mean) of the data in the cluster. In the early stages, *k*-Means algorithm is randomly selected as the centroid *k* objects in D data, then, the distance between the object and the centroid is calculated using euclidian distance. *k*-Means algorithm is iterative increase in the variation in the value of each of each cluster in which the object is placed in the next closest group, calculated from the centre of the cluster (Fig 1b). New cluster centre is determined when the all data has been placed in the nearest cluster.

The process of determining the cluster centre and the placement of data within a cluster is conducted iteratively until the cluster centre value of all clusters are formed and not change anymore (Han et al, 2012).

Various extensions on the ability of *k*-Means algorithm have been done to date. Kumar and Wasan, 2010 recorded three variants of the *k*-Means modified algorithm such as global *k*-Means algorithm (Likas et al, 2003), the efficient *k*-Means algorithm (Zhang et al, 2003) and the X-Means algorithm (Pelleg and Moore, 2000). The increased ability of the *k*-Means algorithm has been carried out with the proposed fast adaptive k-means clustering algorithm (Darken and Moody, 1990, intelligent *k*-Means (Mirkin, 2005), improved genetic *k*-Means algorithm (Guo et al, 2006), intelligent constrained *k*-Means (Amorim, 2008) and the proposed shift of the mean-based initialization on the *k*-Means (Cabria and Gondra, 2012).



(a) Initial Cluster (b) Iteration Process (C) Final Cluster

Fig 1. Object clustering process using *k*-Means algorithm

k-Means algorithm uses spatial data on the current implementation of the new various applications such as fire risk data clustering in the urban area (Lizhi and Aizhu, 2008), identifies clusters of trees using satellite image data (Fan et al, 2010) and the planning of transportation systems conjunction with the determination of the appropriate number and location of service centres of the cassava (Tangkitjaroenmongkol, 2011). Specifically *k*-Mean algorithm implementation for seismic hazard data at this time is still limited and the combination of the algorithm and kriging algorithm to predict the unknown values at specific locations with the aim of preparing spatial zonation is a major contribution of this paper.

3 Methodology

The data used in this study consists of the data peak ground acceleration (PGA), lithology and topography zone data. Data conversion processes have been done for lithology and topography zone from the map into the data with the value of each class based on contribution data at the level of buildings damage. The software used is statistical data processing software package with an additional R Rattle (Williams, 2001) for data clustering and clValid package (Brock et al, 2008) to determine the optimum number of clusters.

The study consisted of four stages, namely (1) data normalization, (2) determining the optimum number of clusters, (3) Data clustering using partitioning *k*-Means algorithm and (4) the data interpolation using kriging algorithm to construct the spatial patterns of building damage hazard area.

1. Data Normalization

Normalization of data using Min-max method that produces a linear transformation of the original data becomes a new data. The Min-Max normalization will be maps a value v of A to v' in the range of new minimum

and maximum value (Han et al, 2012). Min-max normalization formula is shown in equation 1.

$$v = \frac{v - minA}{maxA - minA} * (new_{maxA} - new_{minA}) + new_{minA} \quad (1)$$

In where minA, maxA, new_{maxA} and new_{minA} each is a minimum value and a maximum value of attribute A and the maximum and minimum values on the new scale of attributes A.

2. Determining Optimum Cluster Number

Determination of the optimum number of clusters available data conducted by three approaches, namely (a) Connectivity index, (b) Dunn Index and (c) Silhouette Index.

Tambahkan informasi mengenai Index index yang digunakan

Each index formula as follows:

a. Connectivity Index

$$Conn(C) = \sum_{i=1}^{N} \sum_{j=1}^{L} X_{i,nni(j)} \quad (2)$$

Where nni (j) is the observation nearest neighbor (nearest neighbor) i to j and L as a parameter that determines the number of neighbors that contribute to the measurement of connectivity.

b. Dunn Index

$$D(C) = \frac{\min_{Ck,Cl \in C, Ck \neq Cl} \left(\min_{i \in Ck} \min_{j \in Cl} dist(i,j) \right)}{\max_{Cm \in C} diam(Cm)} \quad (3)$$

Where *diam*(Cm) is the maximum distance between observations in the cluster Cm with an index value between 0 to ∞

c. Silhouette Index

Silhouette index is calculated as the degree of confidence in the clustering process on an observation. Cluster is said to be well-formed if the index value close to 1 and vice versa if the condition index values approaching -1. Silhouette index calculated by the formula:

$$S(i) = \frac{bi - ai}{\max(bi, ai)} \quad (4)$$

Where ai is the average distance between i and all other observations in the same cluster, and bi is the average distance between observations in cluster i with the nearest neighbors. Ai and bi values calculated by the formula:

$$ai = \frac{1}{n(C(i))} \sum_{j \in C(i)} dist(i,j) \quad (5)$$

$$bi = \min_{Ck \in Cl \setminus Ci} \sum_{j \in C(k)} \frac{dist(i,j)}{n(Ck)} \quad (6)$$

Where C (i) is clustering the observation i, dist (i; j) is the distance between the observation i to j, and n (C) is the cardinality of the group C. Silhouette value is in the range of 1 to -1.

3. Data Clustering Using Partitioning *k*-Means Algorithm

k-Means algorithm to put an object into a particular cluster based on the value of the average (mean) nearby. In its simplest form, the algorithm consists of three phases:

a) Determine the number of k cluster and also specify in the center of each cluster (initial cluster center generally is *k* in the first observation or randomly determined).
b) Calculate the distance between each object with each of the cluster center of each object. Insert each object to the cluster based on the closest distance to the center of the corresponding cluster.
c) Calculate the value of the average cluster central for the newly formed cluster.

Repeat step 2 until no longer needed object transfer between clusters. Final determination of an object to a particular cluster is independent of *k* initials were first determined

4. Data Interpolation Using Kriging Algorithm to Construct Spatial Pattern

The next stage is plotting the average value of the results of clustering on each grid using GIS software and the data interpolation with kriging algorithm for the preparation of spatial patterns. Kriging equations composed by a linear combination of:

$$Z(x0) = \sum_{i=1} \omega i(x0) Z(xi) \quad (7)$$

$$\hat{Z}(x_0) = \begin{pmatrix} z_1 \\ \vdots \\ z_n \end{pmatrix}' \begin{pmatrix} c(x_1,x_1) & \cdots & c(x_1,x_n) \\ \vdots & \ddots & \vdots \\ c(x_n,x_1) & \cdots & c(x_n,x_n) \end{pmatrix}^{-1} \begin{pmatrix} c(x_1,x_0) \\ \vdots \\ c(x_n,x_0) \end{pmatrix} \quad (8)$$

4 Result and discussion

4.1 Determination of Optimum Number of Clusters

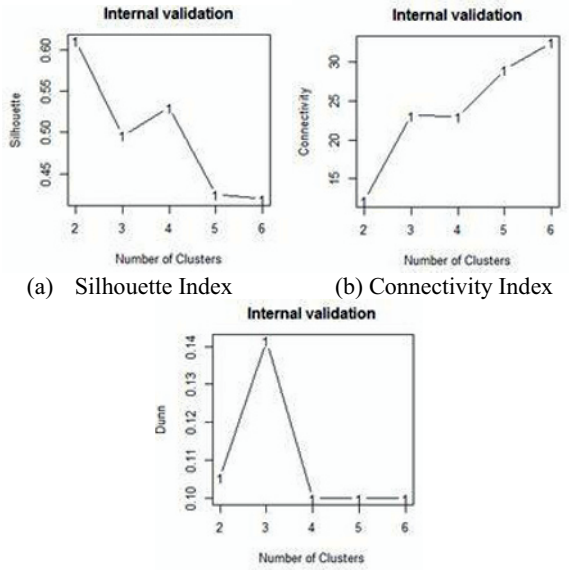
Determination of the optimum number of clusters that implements clustering algorithm, commonly done in the early stages of data mining research as has been done by Brock et al, 2008 and Ahmad et al, 2012. Brock et al, 2008 developed clValid package in R software to analyze the optimum number of clusters using data from Affymetrix microarrays to compare gene expression of mesenchymal cells from two different lineages use a Connectivity index, Dunn and Silhouette with various clustering methods and variations in the number of groups. Concluded that the optimum number of clusters for the data used is 2 (two) with a hierarchical clustering method. Ahmad et al, 2012 in their study using seven (7) kinds of indices to determine the optimum number of clusters for regionalization of rainfall data in peninsular Malaysia. Dunn and Silhouette index used in the study concluded that the optimum group for the data used is 2 (two).

The optimum number of clusters in this study was analyzed using three types of indices as used by Brock et al, 2008, and two of which are used by Ahmad et al,

2012. Determination of the connectivity index based on the value of the smallest index and is different from the use of the Dunn index and Silhouette index which is based on the biggest value index of the overall index value is generated. Using the connectivity index and Silhouette index for three variable of earthquake building damage assessment with the help of statistical data processing software package R with clValid (Brock et al, 2008), concluded that the cluster is ideal for those type of data used are two clusters with the smallest index value of connectivity 12.1679 and the largest value of Silhouette index is 0.6100. The optimum number of different clusters resulting from the use of the Dunn index with the optimum cluster number three with the largest index value of the overall index value generated is 0.1414. Values for the three indices used in the number of group 2 to 6 are as shown in Table 1 and Fig 2.

Table 1. Index value for number of cluster 2 to 6.

Cluster Number	2	3	4	5	6
INDEKS					
Connectivity	12.167	23.128	23.004	29.038	32.534
Dunn	0.105	0.141	0.100	0.100	0.100
Silhouette	0.610	0.496	0.530	0.424	0.419



(a) Silhouette Index (b) Connectivity Index

(c) Dunn Index

Fig 2. Variability of value index in different number of cluster

4.2 Data clustering using *k*-Means algorithm

Early stages of the process of spatial data clustering using partitioning *k*-Means algorithm is to determine the initial cluster centers that subsequently kept updated until the end of the cluster centers generate the final cluster centers for each data variable. Initial cluster centers are used and the resulting final cluster centers for the earthquake building damage hazard data with three variables used in this study are as shown in Table 2.

Using *k*-Means algorithm, the data used 225 divided into 2 (two) clusters based on the distance data relative to the cluster center with the distance between the center of each final cluster is 0.902

Table 2. Initial and final cluster centre.

Data Variable	Initial Cluster Centre		Final Cluster Centre	
	1	2	1	2
X0Norm	.0	1.0	.0	.7
X1Norm	.0	1.0	.0	.7
X2Norm	1.0	.5	.5	.5

A total of 170 data is grouped in cluster 1 (first) and as many as 55 other data is in cluster 2 (two). Distance data on cluster 1 (one) from the cluster center shows a diminishing value at distances further away from the cluster center and most of the compile cluster at a distance of less than 0.1 to 0.5 and in the extreme reduced after 0.5. Distance data on cluster 2 (two) from the cluster centers are mainly concentrated in the range 0.3 - 0.5 the diminishing number at a distance closer or farther away from the center of the cluster (Table 3). Distribution of data for each cluster center in three-dimensional space is, as can be seen in Fig 3

Table 3. Data distance against cluster centre

Distance	Number of data in the Cluster 1 (Totally 170 Data)	Number of data in the Cluster 2 (Totally 55 Data)
< 0.1	41	1
0.1 < D < 0.2	40	7
0.2 < D < 0.3	35	7
0.3 < D < 0.4	29	13
0.4 < D < 0.5	23	17
>0.5	2	7
0.5 < D < 0.6		2
>0.6		1
Total	170	55

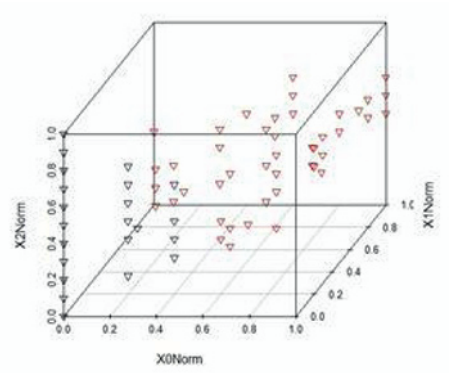


Fig 3. Data distribution in three dimensional space

In addition to the distance to the cluster center, clustering of data in this study was also based on the contribution of each component of the data to the danger of damage to buildings caused by the earthquake. Data with large PGA values, sand lithology with high porosity and low topographic zones and unstable such as seabed class, is classified as a high contribution to the building damage. Instead, the data with a small PGA, clay

lithology with very low porosity and class of high topographic zones and Inland Plain stable class is classified as a low contribution to the building damage

4.3 Zonation of building damage hazard in the city of Banda Aceh

Zonation study of building damage hazard area especialy in the city of Banda Aceh as impact of Sumatra mega earthquake has been conducted by Sengara, 2008; Irwansyah, 2010 ; Irwansyah and Hartati, 2012.

In more detail and using primary data from field survey (Sengara, 2008) constructed a micro-zonation earthquake hazard area in sub Meuraxa, Banda Aceh and divides the study area into three zones: the zone hard soil, medium soil and soft soil zone. Irwansyah, 2010 conducted a study of spatial patterns of building damage using satellite image data with nearest-neighbourhood algorithm. Studies conclusion was that the buildings in the city of Banda Aceh were damaged partially to completely destroy by damage class patterns are relatively parallel to the coastline. Irwansyah and Hartati, 2012 using three variable and processing the cluster using self-organizing map (SOM) and the kriging algorithm, concludes that the City of Banda Aceh can be divided into three classes with the building damage class showing spatial patterns are relatively parallel to the shoreline.

Using the optimum number of clusters and clustering results of data with *k*-Means algorithm as discussed earlier, this study implements kriging algorithm to interpolate the data in order to formulate the spatial pattern (zonation) of building damage hazard area in the city of Banda Aceh. Each cluster data and data cluster spatially is placed back in the location of the data in the center of each grid are total of 224. Subsequent kriging algorithm used to calculate the inter-class values that can be drawn boundaries between classes in order to prepare hazard area zoning of building damage caused by the earthquake. Results show the zoning of two classes of spatial patterns are continuous and relatively parallel to the shoreline and trending southwest - northeast (Figure 4a) with a class of damages hazard are generally located close to the coastal area. Damage hazard class are getting lower towards the inland part the study area. Spatial pattern of the damage hazard area as a resulting from these studies show a similar pattern to the spatial patterns as research Irwansyah and Hartati, 2012 (Fig 4b).

5. Conclusion

Optimum clusters for spatial data hazards caused by earthquake damage to buildings used in this study are two, according to the results of the analysis using the connectivity index and Dunn index.

Using *k*-Means algorithm, the data used in the study were divided into two clusters. A total of 170 data on the first group largely concentrated at a distance of less than 0.1 to 0.5 from the cluster centre, as many as 55 other data on second group largely concentrated at a distance of 0.3 - 0.5 from the cluster centre.

Banda Aceh city and surrounding areas spatially divided into two classes of building damage caused by the earthquake hazard. Spatial patterns of the hazard formed continuously and relatively parallel to the shoreline and trending southwest - northeast. Area with high of building damage hazard are generally located near the coastline and area with low hazard towards to mainland area

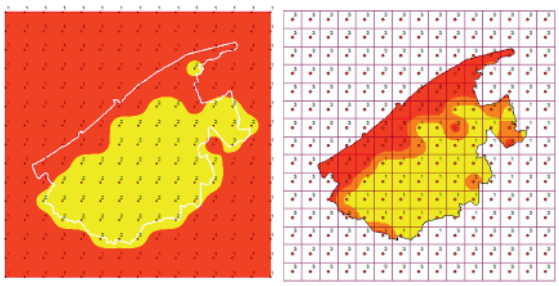


Fig 4. Spatial pattern of building damage hazard in the city of Banda Aceh. (a) Data Clustering using *k*-Means algorithm and zonation using Kriging algorithm and (b) Data clustering using SOM algorithm and dan zonation using dengan Kriging algorithm (Irwansyah dan Hartati, 2012). Red color : area with high of building damage hazard and the yellow area with low of building damage hazard.

References

1. Amorim. R.C. Constrained Intelligent KMeans:Improving Results with Limited Previous Knowledge. The Second International Conf. on Adv Eng Computing and Applications in Sciences (2008)
2. Ahmad.N.H., Othman.I.R and Deni.M. Hierarchical Cluster Approach for Regionalization of Peninsular Malaysia Based on the Precipitation Amount. J. Phys.: Conf. Ser. **423** 012018 (2013)
3. Brock. G., Pihur.V., Datta.S and Datta S. clValid: An R Package for Cluster Validation. Journal for Stastical Software. V **25** Issue 4 (2008)
4. Han.J, Kamber.M, and Tung. A. K. H. Geographic Data Mining and Knowledge Discovery, chapter Spatial Clustering Methods in Data Mining: A Survey, pages 1–29. Taylor and Francis, (2001)
5. Han.J., Kamber.M and Pei.J. Data Mining Concept and Techniques (Third Edition). Morgan Kaufmann-Elsevier (2012)
6. Hartigan.J.A and Wong.M.A. A K-Means Clustering Algorithm-JSTOR J of the Royal Statistical Society. Series C (Applied Statistics), V **28**, No. 1 (1979), pp. 100-108 (1975)
7. Irwansyah. E. Bulding Damage Assessment Using Remote Sensing, Aerial Photograph and GIS Data: Case Study in Banda Aceh after Sumatra Earthquake 2004. Proc The 11th Seminar on Intelligent Tech and Its App-SITIA 2010 V **11** pp.57 (2010)
8. Irwansyah.E and Hartati.S. Zonasi Daerah Bahaya Kerusakan Bangunan Akibat Gempa Menggunakan Algoritma SOM Dan Algoritma Kriging (in Bahasa). Seminar Nasional Teknologi Informasi (2012).

9. Irwansyah E, Winarko E, Z.E Rashid and R.D Bekti . Earthquake Hazard Zonation Using Peak Ground Acceleration (PGA) Approach J. Phys.: Conf. Ser. **423** 012018 (2013)
10. Jain. A.K. Data Clustering: 50 Years Beyond K-Means. Pattern Recognition Letters, 2009
11. Kumar.P and Wasan S.K.. IJCSNS Int J of Comp Science and Network Security, V **10** No.4 (2010)
12. Likas,A., Vlassis, M. & Verbeek, J. The global k-means clustering algorithm, Pattern Recognition, V **36**, p451-461 (2003)
13. MacQueen, J.B. Some Methods for Classification and Analysis of Multivariate Observations. In Proc. of 5th Berkley Symposium on Mathematical Statistics and Probability, Volume I: Statistics, pp. 281–297 (1967)
14. Mirkin, B., Clustering for Data Mining: A Data Discovery Approach, Chapman and Hall/CRC, Boca Raton Fl. USA (2005)
15. Pelleg.D and Moore.A. X-means: Extending K-means with Efficient Estimation of the Number of Clusters, ICML 2000 (2000)
16. Sengara. I.W. Seismic Hazard and Microzonation For A District In Banda Aceh City Post 2004 Great Sumatra Earthquake. The 14th World Conf on Earthquake Eng (2008)
17. Steinbach. M., Karypis. G & Kumar.V. A Comparison of Document Clustering Techniques (2000)
18. Tangkitjaroenmongkol.R,Kaittisin.S & Ongwattanakul.S. Inbound Logistics Cassava Starch Planning With Application of GIS and K-means Clustering Methods in Thailand. 8th Intl Joint Conf on Comp Science and Soft Eng –JCSSE (2011)
19. Tesfamariam.S and Saatcioglu.M. Seismic Vulnerability Assessment of Reinforced Concrete Buildings Using Hierarchical Fuzzy Rule Base Modeling. Earthquake Spectra. V **26** No 1 p235-256 (2010)
20. USGS. Historic World Earthquakes http://earthquake.usgs.gov/earthquakes/world/historical_country.php#indonesia. Last Mod: Nov 01 (2012)
21. Williams G. Data Mining With Rattle and R: The Art of Excavating Data for Knowledge Discovery. Springer (2011)
22. Zhang Y., Mao J. and Xiong Z. An efficient Clustering algorithm, In Proceedings of Second International Conference on Machine Learning and Cybernetics (2003)