Spatial Cluster for Clustering the Influence Factor of Birth and Death Child In Bogor Regency, West Java

Rokhana Dwi Bekti* and Ro’fah Rachmawati

Department of Statistic, School of Computer Science, Bina Nusantara University
**Email: rokhana_db@binus.ac.id*

Abstract. The number of birth and death child is the benchmarks to determine and monitor the health and welfare in Indonesia. It can be used to identify groups of people who have a high mortality risk. Identifying group is important to compare the characteristics of human that have high and low risk. These characteristics can be seen from the factors that influenced it. Furthermore, there are factors which influence of birth and death child, such us economic, health facility, education, and others. The influence factors of every individual are different, but there are similarities some individuals which live close together or in the close locations. It means there was spatial effect. To identify group in this research, clustering is done by spatial cluster method, which is view to considering the influence of the location or the relationship between locations. One of spatial cluster method is Spatial ‘K’luster Analysis by Tree Edge Removal (SKATER). The research was conducted in Bogor Regency, West Java. The goal was to get a cluster of districts based on the factors that influence birth and death child. SKATER build four number of cluster respectively consists of 26, 7, 2, and 5 districts. SKATER has good performance for clustering which include spatial effect. If it compare by other cluster method, K-means has good performance by MANOVA test.

Keywords: spatial, cluster, SKATER
PACS: 02.50.Ng

INTRODUCTION

Birth and death child is one of the benchmarks for assessing the extent of the community welfare achievement in the health development. According to the data from Ministry of Health Indonesia [1], the rank of crude birth rate in Indonesia 2010 in the ASEAN was 5. It was 20 births per 1000 population and the death mortality rate was 30 deaths per 1000 live births. The number of births child in West Java Province 2007 was 822.481 and the number of born child which death was 2.575. In Bogor, West Java Province, there are 88.633 children who are live-born which life and the 210 dead born [2].

In addition, birth and death is used for monitoring the health situation, the input of number population estimation, and to identify groups of people who have a death high risk. Identifying group of death risk is important to compare the characteristics of human that have high and low risk. These characteristics can be seen from the factors that influenced it.

The factors that influence birth and death child are internal factors of parents (age, amount of children, and distance between birth child), economy, sanitation, healthy, education, and health facility. The research by [3] stated that economic, educational, factors sanitation, and health factors were influence the quality of public health, including the birth child. The influence factors of every individual are different, but there are similarities some individuals which live close together or in the close locations. It means there was spatial effect.

The methods for clustering which regard on space and time are spatial cluster. Spatial method is a method to get information of observations influenced by space or location effect [4]. So the spatial cluster was build clustering of region by considering the location of the regions. This method was used because there are many factors which influence of birth and death child, such us economic, health facility, education, and others. The influence factors of every individual are different, but there are similarities some individuals which live close together or in the close locations.

There are some spatial methods to identify clustering, such us Tango’s Index and Kulldorf spatial scan statistics [5], Moran’s I and Geary’s I [6]. These methods were also used in disease cases, such us [7] which was clustering malaria disease in Western Kenya and [8] was clustering the epidemic Cholera. Other method is Spatial ‘K’luster Analysis by Tree Edge Removal (SKATER). The algorithm used was a

4th International Conference on Mathematics and Natural Sciences (ICMNS 2012)
AIP Conf. Proc. 1589, 468-471 (2014); doi: 10.1063/1.4868845

strategy for transforming the regionalization problem into a graph partitioning problem [9].

The clustering of human is important to identify the community welfare achievement in the health development, so this research was perform a spatial cluster by SKATER. The variables used are the factors which influenced birth and death child. The clustering performs on districts in Bogor Regency.

METHODOLOGY

This research was located in Bogor Regency, West Java Province, Indonesia. This region has 40 districts. The data sources are from Central Bureau of Statistics Indonesia and Department of Health year 2008. The map of location was shown in Figure 1. The variables used in clustering were (1) ratio of number of medical personnel and populations (X_1), (2) ratio of number of health center and populations (X_2), and (3) the percentage of pre-prosperous households (X_3).

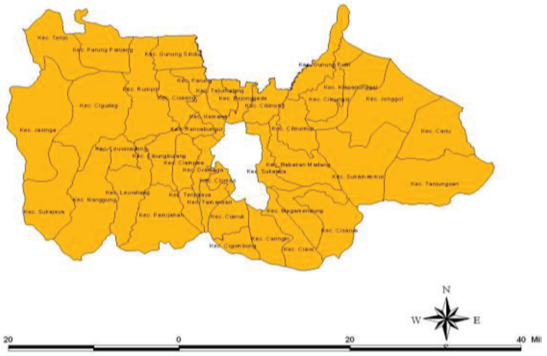

FIGURE 1. 40 Districts in Bogor Regency

The step of analyze are (1) exploration the data, (2) clustering by K-means, (3) clustering by SKATER, and (4) compare the results of K-means and SKATER clustering by MANOVA test. The algorithm of SKATER consist of generate Minimal Spanning Tree (MST), and MST partitioning. This research use R software to compute the analysis.

SKATER Cluster

Spatial 'K'luster Analysis by Tree Edge Removal (SKATER) use algorithm which is a strategy for transforming the regionalization problem into a graph partitioning problem. SKATER is done by two steps [10]. First, it creates a minimal spanning tree (MST) from the graph representation for the neighborhood structure of the geographic entities. The cost of an edge represents the similarity of the entities attributes, defined as the Euclidean squared distance between them. The MST represents a statistical summary of the neighborhood graph based on the entities attributes.

Consider a set of areal spatial objects O with a set of attributes $\{A_1, \ldots, A_n\}$. All objects have an attribute vector $x=(a_1, \ldots, a_n)$ where a_1 is a possible value of the attribute A_1. The topology of the set determines a connectivity graph G=(V, L) with a set of vertices V and a set of edges L. There is an edge connecting vertices v_i and v_j if areas i and j are adjacent. Associate a cost d(i, j) with the edge (v_i, v_j) by measuring the dissimilarity between objects i and j using their attribute vectors x_i and x_j. Euclidean distance for vectors x_i and x_j is

$$d_{ij} = d(x_i, x_j) = \sum_{l=1}^{n} \left(x_{il} - x_{jl}\right)^2 \qquad (1)$$

Building the MST based on Prim's algorithm shown from Jungnickel (1999) in [11]. These algorithms give a connectivity graph for a set of vertices and a set of edges on each location. These algorithms are,

a. Suppose connectivity graph G= (V, L) with a set of vertices (V) and a set of edges (L), the algorithm starts with one T_1 tree.
   Choose any vertex v_i in the complete set of vertices (V), setting T_k=T_1= ($\{v_i\}, \phi$)
b. Find the edge of lowest cost (l') in L that connects any vertex of T_k to another vertex, v_j, belonging to V but not to T_k.
c. Add v_j and l' to the tree T_k, and then creating a new tree T_{k+1}.
d. Repeat step (b) until all vertices have been included in the tree (T_n).

The second step, SKATER performs a recursive partitioning of the MST to get contiguous clusters. It considers explicitly the clusters internal homogeneity. To make a partition of n objects in k regions, it is necessary to remove k – 1 edges from the MST [11]. Each resulting cluster will be a tree. The partitioning algorithm produces a graph G* that contains a set of trees T_1, …, T_n where each tree is connected but has no common edges or vertices with the other trees.

The selected edge performs by sum of the intracluster square deviations, which needs to be minimized:

$$Q(\Pi) = \sum_{l=0}^{k} SSD_i \qquad (2)$$

where Π is a partition of objects into k trees,Q(Π) is a value associated with the quality of a Π partition, and SSD_i is the sum of square deviations in region i.

K-Means Cluster

K-means Cluster is one method of nonhierarchical clustering algorithms. It works by partitioning the data into a user-specified number of clusters [11]. This process iteratively reassigning observations to cluster until some numerical criterion is met. This research set 4 specified numbers.

MANOVA

Multivariate analysis of variance (MANOVA) is simply an ANOVA with several dependent variables. It tests for the difference in two or more vectors of means [11]. This test use the hypothesis

H_0: there is no difference in the average
H_1: there is a difference in the average

The conclusion is rejecting Ho if F value higher than F table with v_1 and v_2 degrees of freedom. In this research, MANOVA was used for compare cluster methods based on test of mean in each cluster. The high F value and small P value shows that there were different mean in some cluster.

RESULTS

Central Bureau of Statistics show that Bogor Regency has total populations 4.340.520 in 2008 and 4.477.344 in 2009. This number was continuous ascending since five years ago. Department of health shown that there are 94.665 children born which life and the 270 children born was death in 2008 [1]. This number was higher than one year ago. This condition was the reason why need to study about identify the factors which influence on birth and death child. That identify was done by clustering such as spatial cluster SKATER method. Then compare this method with K-means which not include the spatial effect.

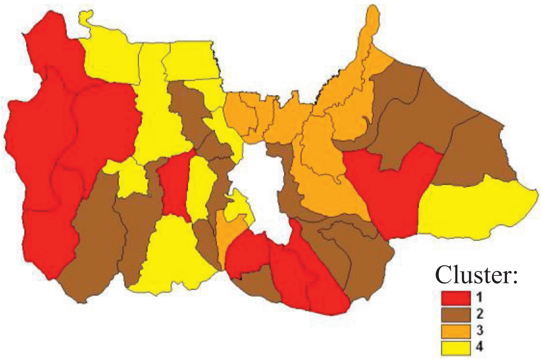

FIGURE 2. Clustering by K-means

The first clustering method performed by K-means. The result was presented in Figure 2. Three variable used for clustering are ratio of number of medical personnel and populations (X_1), ratio of number of health center and populations (X_2), and the percentage of pre-prosperous households (X_3). Clustering was built from four main groups or cluster. The first group consists of 9 districts. The second group was 13 districts. The third group is 8 districts, and the fourth group was 10 districts. From the figure, it is seen that the results of clustering spread evenly across the entire area Bogor Regency. The pattern of spatial relations cannot be seen through this figure. Spatial patterns can be represented by the same characteristics in the districts that are close together.

Further identification was use spatial cluster SKATER. The first step is to establish a Minimum Spanning Tree (MST) Generation. The results are as follows in Figure 3. It shows the connectivity graph of 40 districts. As in Prim's algorithm, it builds 40-1=39 edge. The first edge consists of vertices in location 5 and 2 which build tree (T_1) and the cost was 12,091. The second edge consists of vertices in location 2 and 1 which build tree (T_2) and the cost was 3,401. The last edge consist of vertices in location 18 and 19 which build tree (T_{40}) and the cost was 12,951.

The second step was determining the MST Partitioning (Figure 4). This step is build number of group or cluster. It was shown by the line in the map which has different colors. Example, the first cluster was location (district) 36 and 37 which was shown by blue line. The second cluster was location (district) 38,33,32,31, and 30 which was shown by purple line. In the clustering process, it determined that there were four number of cluster. The first group consists of 26 districts. The second group was 7 districts. The third and four groups were 2 districts and 5 districts.

Figure 5 has shown that some districts which in the same cluster were close together. It means that spatial pattern can be seen through this figure. There were 26 districts in the first group and close together among districts. There were located on southern of Bogor Regency. Other clusters were having the some interpretation. All of the districts in the some cluster were closed together and on the north of Bogor Regency. This clustering showed the characteristic distribution of the birth and death ratio from population of each district, the ratio of health workers, hospitals and the percentage of poor families in Bogor.

The results clustering by K-means and SKATER were different. The pattern of spatial relations cannot be seen through K-means. But, the spatial pattern can be seen by SKATER. To compare the results, this research use F test in MANOVA (see Table 1). MANOVA can be used for compare these two methods based on test of mean in each cluster. The high F value and small P value shows that there were different mean in some cluster. The results in Table 1 show that K-means has good performance than SKATER. It was shown by F value in K-means which higher than in SKATER. Although this test conclude that K-means has good performance, but it cannot perform spatial pattern. It can be seen from Figure 1. SKATER has good performance when look on spatial pattern in Figure 5.

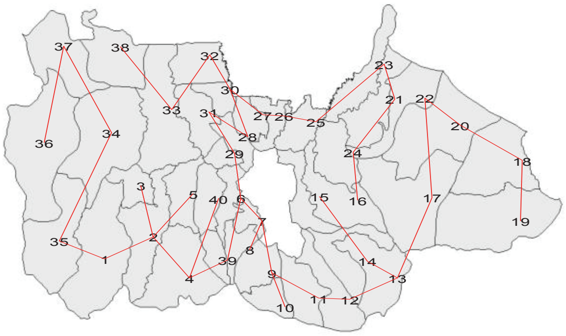

FIGURE 3. Connectivity Graph in process of Minimum Spanning Tree (MST) Generation

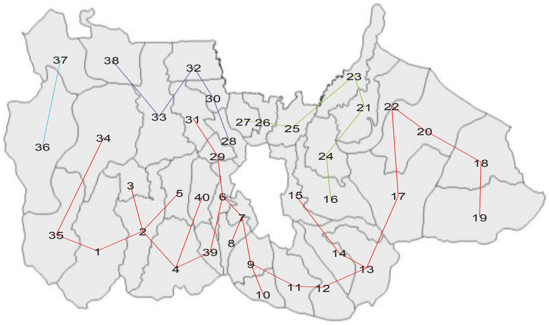

FIGURE 4. Minimum Spanning Tree (MST) Partition

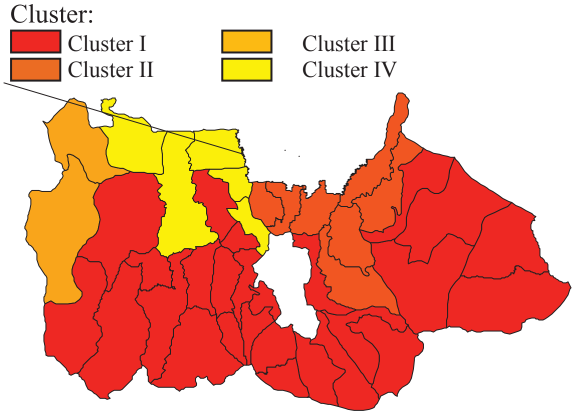

FIGURE 5. The Results of Clustering by SKATER

TABLE 1. MANOVA in K-means and SKATER Clustering

Method	F value	P-value
K-means	15.014	0.000
SKATER	1.6541	0.1942

CONCLUSSION

The result of clustering using K-means were the first group consists of 9 districts, the second 13 districts, the third 8 districts, and the fourth 10 districts. SKATER was built the first group which consists of 26 districts, the second 7 districts, the third 2 districts and the fourth 5 districts. All districts in the some group are close together, such us the first group which located on southern of Bogor Regency. It means that this method perform the clustering by spatial influence. K-means has good performance, but it cannot perform spatial pattern. SKATER has good performance when look on spatial pattern.

REFERENCES

1. [KemenKes RI], *Profil Kesehatan Indonesia 2010*, Jakarta:Kementerian Kesehatan Republik Indonesia, 2011.
2. [Dinkes Jabar], *Profil Kesehatan Provinsi Jawa Barat 2007*, Bandung : Dinas Kesehatan Provinsi Jawa Barat, 2008.
3. D. Winarno, *Analisis Angka Kematian Bayi Di Jawa Timur dengan Pendekatan Model Regresi Spasial*, [Tesis], Surabaya: Program Sarjana Jurusan Statistika ITS, 2009.
4. L. Anselin, *Spatial Econometrics: Methods and Models*, Ist Edn., Netherlands : Kluwer Academic Publishers, 1988.
5. T. Tango, *Statistical Methods for Disease Clustering*, USA : Springer, 2010.
6. J. Lee and D.W.S. Wong, . *Statistical Analysis with Arcview GIS*, Ist Edn., New York : John Wiley and Sons, 2001, pp: 208.
7. S. Brooker, S. Clarke, JK. Njagi, S. Polack, B. Mugo, B. Estamble, E. Muchiri, P. Magnussen, J. Cox, Spatial clustering of malaria and associated risk factors during an epidemic in a highland area of western Kenya, *Tropical Medicine and International Health*, 2004. **9**:757-766.
8. D.R. Moreno, M. Pascual, M. Emch, M. Yunus, Spatial Clustering in the Spatio-Temporal Dynamics of Endemic Cholera, BMC Infectious Diseases, 2010, pp. 10:51.
9. R.M.Assuncao,M.C.Neves, G.Camaras and C.DA.C. Freitas, Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees, *International Journal of Geographical Information Science*, 2006, 20:7, pp 797-811
10. A.I. Reis, G. Camara, R. Assuncao, AMVM. Monteiro, Data-Aware Clustering for Geosensor Networks Data Collection, Anais XIII Simpósio Brasileiro de Sensoriamento Remoto, Florianópolis, Brasil, 2007, pp. 6059-6066.
11. Hair, J.F., Anderson, R.E., Tatham, R.L. dan W.G. Black, Multivariate Data Analisis, fifth edition, Prentice-Hall International, Inc, 1998.