



www.cafetinnova.org

Theme of the Paper: **REMOTE SENSING AND  
GIS APPLICATION**

Jointly organizing by

Andalas University, Padang, INDONESIA  
Cafet Innova Technical Society, Hyderabad, INDIA

**International  
Conference  
on Earth  
Sciences  
and Engineering**

ICEE 2017, 29<sup>th</sup>-31<sup>st</sup>  
August, 2017

# Clustering of Social Conditions in Batam, Indonesia Using K-Means Algorithm and Geographic Information System

**DWI ELY KURNIAWAN<sup>1</sup>, AGUS FATULLOH<sup>2</sup>**

Department of Informatics Engineering, Batam State Polytechnic, Batam, Indonesia

Email: [dwialikhs@polibatam.ac.id](mailto:dwialikhs@polibatam.ac.id) , [agusf@polibatam.ac.id](mailto:agusf@polibatam.ac.id)

---

**Abstract:** Batam is an archipelago that has different social characteristics of each sub-district. Social problems should be a concern and responsibility of both the public and government. For the programs launched by the government in accordance targeted and it is necessary to determine the priorities of the social problems that arise in each region. This study uses k-means algorithm to classify social problems in the form of clusters. Each group has the same characteristics in the group while the other cluster to form the next group according to their characteristics. The results of the study, a k-means algorithm determines the group and map the distribution of social issues into the Geographic Information System (GIS) to facilitate the social information to know each district and identify priorities for help to the government program on target.

**Keywords:** clustering, social issues, GIS

---

## 1. Introduction

Batam Island is one of the largest islands in the Riau Islands. Every year, the population growth rate is always increased. The increase in the population growth rate due to the many people from outside who came to the island of Batam to look for a job. According to statistics of Batam City area, the total population in 2014 increased in the year 2015 as many as 6659 of 1,030,528 inhabitants, namely the soul to be 1,037,187 inhabitants. These increases can cause various social problems such as poverty, education, labor, marriage, and divorce as well as other social issues that would adversely affect the socio-economic condition of the people when it is not sought.

Welfare is the ideals and shared expectations outlined in the first UUD'45 and Pancasila. Government and society should be involved in the development and make improvements to the common interest. For the programs launched by the government in accordance targeted and it is necessary to determine the priorities of the social problems that arise in each region. Therefore there is a need to know the socio-economic conditions in each district to facilitate the decision of the government, private, and so on.

One way to do that is by doing clustering. Clustering is a data mining method used to classify the objects that have

a high level of similarity within a cluster, where each group has a high degree of difference with other groups. Similar observations, analysis of land use mining venture using statistical methods K-Means Clustering-based geographic information system [9], the mapping of areas prone to crime in the jurisdiction Poltabes Semarang in 2013 by using the clustering method [15], and identifying the spread of earthquake clustering [16].

Based on these studies, this study apply k-means clustering method to facilitate stakeholders, government or public institutions (CSR) can help define and map out priorities and program guidance so precise targets based on social issues that arise in each district of Batam. Results clustering of social conditions is processed and presented in the form of a Geographic Information System (GIS).

## 2. Related Study

### 2.1. Social conditions

Social problems are symptoms or phenomena that appear in the reality of social life. Social problem as a condition that is not desired by most citizens. Factors causing social problems is essentially an inseparable part of human life. This is because social problems manifest as a result of human culture itself and the result of the relationship with other humans. Among the causes of social problems is measured by: (1) a lack of compatibility between social values with social action; (2) the source of social

problems are the result of a social phenomenon in society; (3) the parties to establish a social phenomenon depends on the characteristics of the society; (4) the manifest social problems and latent social problems; (5) the attention of the public and social problems; (6) the value system and the improvement of a social problem. Social problems arise due to the social inequality between the values in society with the existing reality. The social imbalance occurs the striking difference between the one with the other.

Social problems can generally be categorized into five types of factors such as (1) economic factors, including poverty and unemployment (2) cultural factors, namely the settlement of issues between the family and the community as the divorce rate of street children and juvenile delinquency, (3) biological factors (public health), which resulted in the spread of infectious diseases, food poisoning and so forth, (4) factors including education enrollment rates, illiteracy and graduates, (5) population factors such as the number and distribution of the population.

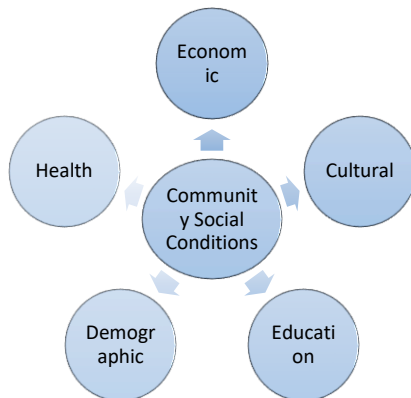


Figure 1. Social Indicators

Indonesia itself there are many social problems are not being resolved, one of which is the problem of poverty. Based on data from the Central Bureau of Statistics (BPS) September 2015, the number of poor people reached 28.51 million (11.13%). The figure indicates that poverty alleviation programs have not successfully overcome the problem of poverty. Basically, poverty reduction programs tend to focus their efforts on the distribution of social assistance for the poor who may not be right on target.

## 2.2. K-Means Clustering Algorithm

K-Means clustering algorithm is the most popular. This algorithm is compiled on the basis of a simple idea. There was initially determined how the cluster to be formed. Any object or the first element in the cluster can be selected to serve as the center (centroid point) cluster.

K-Means algorithm is a non-hierarchical clustering algorithm partitioning the data into the form of one or more cluster or group. This method of partitioning the data into clusters so that data which have the same characteristics are grouped into the same cluster and data that have different characteristics grouped into clusters others.

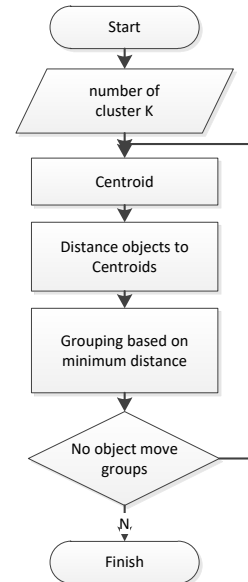


Figure 2. K-Means Clustering Algorithm

Steps Algorithm K-Means, among others:

- 1) Determine the number of clusters (k) and allocate data into cluster
- 2) Determining the centroid randomly
- 3) Calculate the distance of existing data to the centroid
- 4) Segment data is based on the minimum distance to the centroid
- 5) Repeat step three until the position data is not changed.

Cluster grouping process is performed by calculating the shortest distance from the centroid point data. Minkowski distance used to calculate the distance between the data with the following formula [4].

$$d(i,j) = \sqrt[q]{|x_{i1}-x_{j1}|^q + |x_{i2}-x_{j2}|^q + \dots + |x_{ip}-x_{jp}|^q} \quad (1)$$

Description:

- q = 1 to calculate the distance Manhattan
- q = 2 to calculate the Euclidean distance
- q = infinity to calculate the distance Chebychev
- $X_i, X_j$  = data to be calculated the distance
- p = dimension of the data

## 2.3. Geographic Information Systems

GIS is an organized collection of computer hardware, software, geographic data, and personnel designed to efficiently acquire, store, update, manipulate, analyze and display all forms of reference geographic information [6]. GIS is able to assist in solving the problem by displaying the data in a manner that is easily understood and the results are easily disseminated. The software can be used to digitize the spatial data can use Quantum GIS. Digitization in the form of points, lines, and polygons, to display data in the social conditions of each district.

### 3. Research Methods

#### 3.1. Data Collection

Needs analysis starts from the excavation data needs. Excavation social data derived from primary data obtained by conducting surveys and interviews to 12 districts and secondary data obtained from the Department of Social Welfare and the Central Bureau of Statistics (BPS) Batam and related research. Such data must be adjusted to the existing data in each district based data support with social welfare problems (PMKS data) in Batam. Data have been obtained pre-processing is then performed in advance so that the data ready to be processed.

#### 3.2. Processing Data

After collecting the data, the next stage of data processing. Before the processed data pre-processing is done in advance. Pre-processing is done to clean up the data outliers, data errors, missing data and inconsistent data, or contain different values. Data preprocessing is essential to produce good quality mining. Several stages were carried out before the data mining process.

- 1) Data cleaning process: data cleansing is do with regard to detecting, removing errors and inconsistent data to improve data quality. Data errors can include incomplete data, spelling errors, duplicate data. At this stage, the reduction of incomplete data and attributes that are not needed, complete the missing data and correct the data inconsistent.
- 2) Data selection: choosing of data social, which represents the social problems that arise in each district.

**Table 1: Data selection**

District	X1	X2	X3	X4	X5
Belakang Padang	280	21836	13	1978	37
Bulang	82	11625	8	683	42
Galang	89	16585	9	1124	70
Sei Beduk	445	89648	78	1086	31
Nongsa	467	59844	158	1670	57
Sekupang	845	127288	58	3220	26
Lubuk Baja	1134	93868	270	4766	80
Batu Ampar	515	76212	219	1963	15

Batam Kota	1288	155077	133	3697	67
Sagulung	1277	163750	85	2200	57
Batu Aji	552	116153	60	974	29
Bengkong	966	105301	145	2671	55

- 3) Data transformation: transformation of transform complex data does not remove content, making it easier on though. Data that type of character is initialized in advance in the form of numbers. Transformation of data by normalizing. Normalization is standardizing the data to a variable of all the same attributes. Normalization Z-Score is the normalization method based on mean (average value) and standard deviation (standard deviation) of data [4].

$$v' = \frac{v - \text{mean}_A}{\text{stand\_dev}_A} \quad (2)$$

**Table 2: Data Normalization**

District	X1	X2	X3	X4	X5
Belakang Padang	0.8844	1.2554	1.07003	0.15544	0.50537
Bulang	1.3433	1.4538	1.12948	1.20753	0.25682
Galang	1.3271	1.3574	1.11759	0.84925	216
Sei Beduk	0.5021	0.97051	0.29723	0.88012	0.80362
Nongsa	0.4511	0.5167	0.154	0.40567	5164
Sekupang	0.4248	0.7940	0.53501	0.85358	1.05217
Lubuk Baja	1.0945	0.1445	1.98551	2.10959	1.63211
Batu Ampar	0.3398	0.1986	0.888	0.16763	1.59897
Batam Kota	1.4514	1.3340	0.35667	1.24111	0.98589
Sagulung	1.4259	1.5026	0.21400	0.02491	0.48880
Batu Aji	0.2541	0.5776	0.51123	0.97111	0.90304
Bengkong	0.7052	0.3667	0.49935	0.40756	0.38938

Furthermore, integrating the data to be processed. Once the data is ready, then do clustering.

### 4. Results and Implementation

Social conditions used in this study is to capture data as follows social problems.

- 1) Economic factors of society, to describe the state of economy of a region by taking the poverty data for each sub-district of Batam.
- 2) Cultural factors, to define the circumstances and customs of the people in the resolution of problems in the family and social environment. Cultural factors take data on the number of divorces in each sub-district of Batam.
- 3) Educational factors, to describe the state of participation and minimal graduates to seek employment in Batam. Graduates minimal taken is the number of high school level in each sub-district of Batam.
- 4) The population factor, to define the number, composition and distribution of the population in each District of the city of Batam.
- 5) Biological factors (public health), to describe the human health condition of a disease that may occur in each sub-district of Batam.

In determining the most appropriate number of clusters is then performed simulations with cluster number 2, 3 and 4 to be see which form the most appropriate cluster.

**Table 3:** Cluster of results

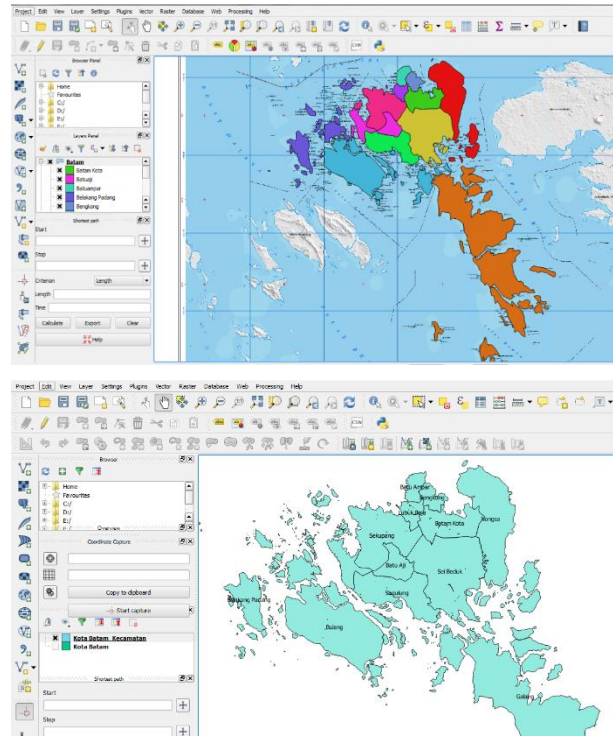
District	Cluster
Lubuk Baja	cluster0
Batam Kota	cluster0
Sagulung	cluster0
Bengkong	cluster0
Sei Beduk	cluster1
Nongsa	cluster1
Batu Ampar	cluster1
Sekupang	cluster2
Batu Aji	cluster2
Belakang Padang	cluster3
Bulang	cluster3
Galang	cluster3

Based on the analysis of linkages between clusters with attributes can be conclude that:

- 1) Cluster 0 Average experienced many problems in economics and health (includes District Lubuk Baja, Batam City, Sagulung, and Bengkong)
- 2) Cluster 1 on average had problems in the field of education and culture (includes District Sei Beduk, Nongsa and Batu Ampar)
- 3) Cluster 2 did not experience the problems of social conditions (includes the District Sekupang and Batu Aji)
- 4) Cluster 3 Average experienced problems in health (includes District Belakang Padang, Bulang, Galang)

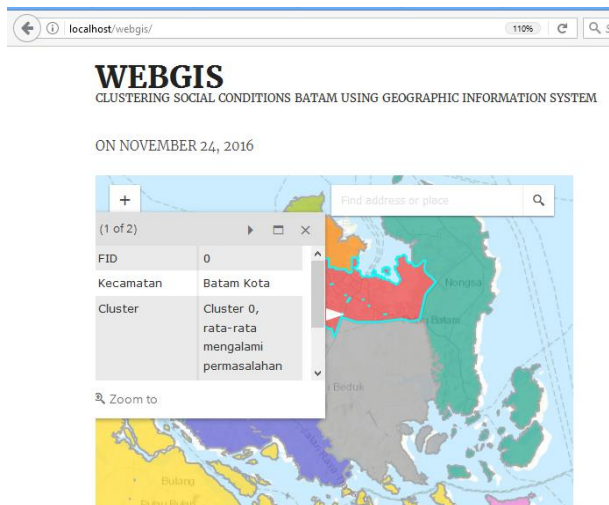
Cluster 0 experiencing economic problems caused by unemployment, street children and social inequality (poverty), while the health factor, means infrastructures exist but still can't meet the eligibility rate due to the number of people that a lot in this district. Cluster 1 is an area that needs to get the attention of educational and cultural programs, where the level of problem-solving in a family environment that divorce is high compared with other districts. This is possible due to the low educational enrollment visible from graduates. Cluster 2 area equitable means good enough. Social problems of these four factors evenly, this is due to existing infrastructure in these districts are adequate and support the social needs. Cluster 3 is the area that needs medical attention because health facilities are inadequate. Cluster 3 is an archipelago far from the city access.

Digitizing polygons (areas) using Quantum GIS conducted by area districts. Figure 3 shows the 12 districts that have been labeled.



**Figure 3.** Results of Digitizing City District Batam

Results digitized maps of Batam using Quantum GIS form that contains a shape file (geometry feature data, index and attributes). The file is used to display in the form of Web GIS. Visualization map using Web GIS based districts to provide social conditions result cluster information from each district.



**Figure 4.** Results of Clustering Web GIS

When the cursor is directed at the spatial districts it will display the cluster information.

## 5. Conclusion

Based on the results, the design can be summarized as follows.

- 1) Social conditions in this study reveal general economic factors, education, population, health and culture.
- 2) Algorithm k-means capable of performing clustering of social data with 4 cluster including Cluster 0 undergone many economic problems and health (includes District Lubuk Baja, Batam City, Sagulung and Bengkong), Cluster 1 on the average, the problems of education and culture (includes District Sei Beduk, Nongsa and Batu Ampar), Cluster 2 did not experience any problems (includes District Sekupang and Batu Aji), Cluster 3 on average experiencing health problems (including Belakang Padang, Bulang, Galang)
- 3) The clustering results are presented in the form of Web GIS so enables users to know the social information.

## 5. Acknowledgment

Author would like to thank for the Centre for Research and Community Service (P2M) Batam State Polytechnic, in funding this research.

## References

- [1] P. Wright, Knowledge Discovery In Databases: Tools and Techniques, available <http://www.acm.org/crossroads/xrds5-2/kdd.html#11>
- [2] P. Berkhin, Survey on clustering data mining techniques, [http://www.ee.ucr.edu/~barth/EE242/clustering\\_survey.pdf](http://www.ee.ucr.edu/~barth/EE242/clustering_survey.pdf)
- [3] H. Garcia-Molina, J.D. Ullman, J. Widom. 2002. Database systems the complete book, International edition. New Jersey, Prentice Hall.
- [4] PN. Tan, M. Steinbach, V. Kumar. Data Mining Cluster Analysis: Basic Concepts and Algorithms. [www-users.cs.umn.edu/~kumar/dmbook/-6k](http://www-users.cs.umn.edu/~kumar/dmbook/-6k).
- [5] G. William. Data Mining Cluster, available [http://datamining.anu.edu.au/student/math3346\\_2005/050809-maths3346-clusters-2x2.pdf](http://datamining.anu.edu.au/student/math3346_2005/050809-maths3346-clusters-2x2.pdf)
- [6] ESRI. 1989. Users Guide ARC/INFO Vol. 2. Common References. The Geographic Information System Software. Environmental System Research Institute, Inc. Redlands, California
- [7] Patrick, O. and Reno, A. 2012. Basic ArcGIS10 Tutorial. Bappeda NTB.
- [8] Suryana. 2011. K-Means Clustering Statistics Method on the Analysis of Land Usage of Mine Based Geographic Information System. Jurnal Statistika, Vol.11 No.1 Hal 7-20 Mei 2011.
- [9] E. Rivani. 2010. Application of K-Means Cluster for Grouping of Provinces Based on Rice, Corn, Soybean, and Green Bean Production Year 2009. Jurnal Mat Stat Vol.10 No.2 July 2010: 122-134
- [10] Z. Zabad. 2015. Social Issues. Available <https://www.scribd.com/doc/259071936/permasalahan-sosial-pdf>
- [11] L. Danem. 2013. Analysis of Population Issues in Batam City and the Solutions.
- [12] Central Bureau of Statistics of Batam City. 2014. Batam In Figures 2014.
- [13] Central Board of Statistics of Riau Islands. 2014. Riau Islands In Figures 2014.
- [14] G.Y. Hilman, B. Sasmito, A.P. Wijaya. 2015. Mapping Areas Prone to Crime in the Law Area of Poltabes of Semarang in 2013 by Using Clustering Method. Journal of Undip Geodesy. Vol.4, No.1, 2015
- [15] R. Pebria, B. Heru, I. Sugihartono. 2011. Identification of Earthquake Spread in Indonesia by Clustering Method. Spectra: Journal of Physics and its Applications, Vol. XII No.2, December 2011