Jurnal Ilmu Komputer dan Informasi (Journal of a Science and Information). 11/1 (2018), 10-16
DOI: http://dx.doi.org/10.21609/jiki.v11i1.580

IMPROVEMENT METHOD OF FUZZY GEOGRAPHICALLY WEIGHTED CLUSTERING USING GRAVITATIONAL SEARCH ALGORITHM

Imam Habib Pamungkas[1] and Setia Pramana[2,3]

[1]BPS Statistics Indonesia, Jl. Dr. Sutomo 6-8, Jakarta, 10710 Indonesia
[2]Institute of Statistic Jakarta, Jl. Otto Iskandardinata No. 64C, Jakarta, 13330, Indonesia
[3]Medical Epidemiology and Biostatistics Department, Karolinska Instutet, SE-171 77 Stockholm, Sweden

E-mail: setia.pramana@stis.ac.id

Abstract

Geo-demographic analysis (GDA) is a useful method to analyze information based on location, utilizing several spatial analysis explicitly. One of the most efficient and commonly used method is Fuzzy Geographically Weighted Clustering (FGWC). However, it has a limitation in obtaining local optimal solution in the centroid initialization. A novel approach integrating Gravitational Search Algorithm (GSA) with FGWC is proposed to obtain global optimal solution leading to better cluster quality. Several cluster validity indexes are used to compare the proposed methods with the FGWC using other optimization approaches. The study shows that the hybrid method FGWC-GSA provides better cluster quality. Furthermore, the method has been implemented in R package spatialClust.

Keywords: *Clustering, Fuzzy Geographically Weighted Clustering (FGWC), Gravitational Search Algorithm (GSA)*

Abstrak

Analisis geo-demografi (GDA) adalah metode yang berguna untuk menganalisis informasi berdasarkan lokasi, dengan memanfaatkan beberapa analisis spasial secara eksplisit. Salah satu metode yang paling efisien dan umum digunakan adalah Fuzzy Geographically Weighted Clustering (FGWC). Namun, ia memiliki keterbatasan dalam mendapatkan solusi optimal lokal pada inisialisasi centroid. Pendekatan baru yang mengintegrasikan Algoritma Pencarian Gravitasi (GSA) dengan FGWC diusulkan untuk mendapatkan solusi optimal global yang mengarah pada kualitas cluster yang lebih baik. Beberapa indeks validitas cluster digunakan untuk membandingkan metode yang diusulkan dengan FGWC menggunakan pendekatan optimasi lainnya. Studi tersebut menunjukkan bahwa metode hibrida FGWC-GSA memberikan kualitas cluster yang lebih baik. Selanjutnya metode tersebut telah diimplementasikan pada paket R spasialClust.

Kata Kunci: *Clustering, Fuzzy Geographically Weighted Clustering (FGWC), Algoritma Pencarian Gravitasi (GSA)*

1. Introduction

Nowadays geographical data are available and easy to be accessed and getting more attention to be included in the analysis and commonly used to observe people behavior based on their location. Geo-demographic analysis (GDA) is the analysis of spatially geo-demographic and lifestyle data [1]. Geo-demographic analysis explores information based on location, utilizing several spatial analyses explicitly.

Geo-demographic analysis often uses clustering techniques to classify the geo-demographic data into groups, making the data more manageable for analysis purposes [2].

GDA relay on two assumptions: (1) two individuals who live in the same area are more likely to have similar characteristics than individuals selected at random, and (2) two areas can be characterized in terms of their population, using demographics and other measures. Based on these two principles, clustering can be applied to group geo-demographic data and lead to meaningful results [1].

In GDA, fuzzy clustering is commonly used with different approaches such as Bezdek's Fuzzy C-means Clustering (FCM), Gustafson-Kessel, Neighborhood Effect, and Fuzzy Geographically Weighted Clustering (FGWC). FCM is the most popular clustering method because it is easy to use and efficient [2]. FGWC was proposed to

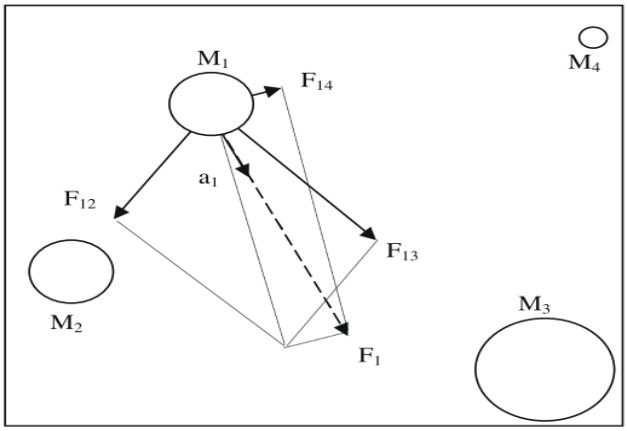

Figure 1. Basic Concept of Object Interaction in GSA [6]

improve FCM on handling spatial data with incorporated geographic and neighborhood data [3]. FGWC was inspired by a hypothesis statement, if we incorporated neighborhood effect to fuzzy clustering, the result will be geographically aware [3].

Similar to the FCM method, FGWC also has limitation in initial phase. The random cluster centroid initialization makes FGWC easily trapped in local optimal solution that effect the cluster quality. Several attempts have been done by using different optimization approach, for example, Artificial Bee Colony (ABC), Particle Swarm Optimization (PSO), and Simulated Annealing (SA) [1].

Gravitational Search Algorithm (GSA) is new optimization algorithm which focus in obtaining a global solution. To improve the cluster quality of FGWC, this research aims to integrate the Gravitational Search Algorithm to avoid FGWC falling in local optimal solution providing better clustering result.

Theoritical Background

Fuzzy C-Means (FCM)

Fuzzy C-Means is one of the most popular clustering algorithms aimed to minimize the following objective function based on membership value of the object and distance of the object to the cluster centroids [4]:

$$J_m(\mathbf{U},\mathbf{V})=\sum_{k=1}^{N}\sum_{i=1}^{c}(\mu_{ik})^m\left\|y_k-v_i\right\|^2 \qquad (1)$$

where "y"_"k" is the k-th observation, c is the initial cluster number, m is degree of the fuzziness, "U" is membership matrix that contains membership degree ("μ"_"ik"), "μ"_"ik" membership degree between the i-th and cluster c, "V" is centroid matrix that contains value of cluster centroid, and "v"_"i" is cluster centroid of cluster c.

Membership degree of each object in FCM is changing during the iteration by using the following function:

$$\mu_{ik}=\left(\sum_{j=1}^{c}\left(\frac{d_{ik}}{djk}\right)^{\frac{2}{m-1}}\right)^{-1};\ 1\le k\le N;\ 1\le i\le , \qquad (2)$$

where "d" is the Euclidean distance between data and cluster centroid. The cluster centroid is defined as follows:

$$v_i=\frac{\sum_{k=1}^{N}(\mu_{ik})^m y_k}{\sum_{k=1}^{N}(\mu_{ik})^m};\ 1\le i\le c \qquad (3)$$

Fuzzy Geographically Weighted Clustering (FGWC)

FGWC proposed by Mason and Jacobson is an extension of version of FCM that more geographically aware [3]. This algorithm takes into account basic spatial interaction effect into the model. The adaptation of spatial effect is performed in each iteration of the following membership matrix calculation:

$$\mu'_i = \alpha\times\mu_i+\beta\times\frac{1}{A}\sum_{j}^{n}w_{ij}\times\mu_i, \qquad (4)$$

where 〖 μ '〗 _i is membership value of area i-th, μ _i is old membership value before incorporating spatial effect, and A is scale value to ensure that sum of membership matrix equal 1. The parameters α and β control the membership proportion after and before weighting a+β=1.

$$w_{ij} = \frac{\left(p_i \times p_j\right)^b}{z_{ij}^a} \quad (5)$$

where p_i and p_j are number of population of area i and j, respectively, and z_ij is the distance between the two areas. The other two parameters, a and β tune the effect of the distance and population on the weight and are defined by the users.

Gravitational Search Algorithm (GSA)

GSA is one of population based algorithm [5] developed by [6]. The aim of this algorithm is improving exploration and exploitation of the population based algorithm to reach optimal solution. GSA is naturally inspired by law of motion and Newtonian gravity.

Every object in GSA is called agent and the capability of each agent measured by his mass. Each agent in GSA will interact based on law of gravity. Agent with small capability will move to agent with large capability.

Figure 1 shows that agent M1 is affected by agents M2, M3 and M4. According to the law of gravity, M1 has a resultant force which will make it move towards agent M3. Agent M2, M3, M4 also has resultant fore to each other.

The first step in GSA is randomly generate initial N solutions with m dimension. The agent position is represented as follows:

$$X_i = (X_{i1}, \dots, X_{id}, \dots, X_{im}), \quad (6)$$

In each iteration, the following total force in each agent (F) is evaluated:

$$F_{ij}{}^d(t) = G(t)\frac{M_i(t)M_j(t)}{R_{ij}(t)}\left(x_i{}^d(t) - x_i{}^d(t)\right), \quad (7)$$

$$F_i{}^d(t) = \sum_{j=1, j\neq i}^{N} rand_i F_{ij}{}^d(t), \quad (8)$$

where ⟦x_i⟧ ^d represents the agent position, (t) is gravitation constant at t, M_i (t) is mass of agent i, and R_ij (t) is the euclidean distance between agent.

$$R_{ij}(t) = \left\|X_i(t), X_j(t)\right\|_2, \quad (9)$$

(t) is updated in each iteration using the following function

$$G(t) = G(G_0, t), \quad (10)$$

where G_0is gravity constant.
The agent mass M_i (t) is defined as follows:

$$m_i(t) = \frac{fit_i(t) - worst(t)}{best(t) - worst(t)}, \quad (11)$$

$$M_i(t) = \frac{m_i(t)}{\sum_{j=1}^{N} m_j(t)}, \quad (12)$$

⟦fit⟧ _i (t) is current fitness value from the solution. The best and worst determined by fitness value. There are two minimization functions to get best and worst:

$$best(t) = \min_{j\in\{1\dots N\}} fit_j(t), \quad (13)$$

$$worst(t) = \max_{j\in\{1\dots N\}} fit_j(t). \quad (14)$$

Whereas the maximization functions:

$$best(t) = \max_{j\in\{1\dots N\}} fit_j(t), \quad (15)$$

$$worst(t) = \min_{j\in\{1\dots N\}} fit_j(t). \quad (16)$$

The acceleration (a) and velocity (v) each agent are defined:

$$a_i^d(t) = \frac{F_i^d(t)}{M_{ii}(t)}, \quad (17)$$

$$v_i^d(t+1) = rand_i \times v_i^d(t) \times a_i^d. \quad (18)$$

The last step is updating the position of each agent x.

$$x_i^d(t+1) = x_i^d(t) + v_i^d(t+1). \quad (19)$$

Repeat step until maximum iteration or reach stopping criterion.

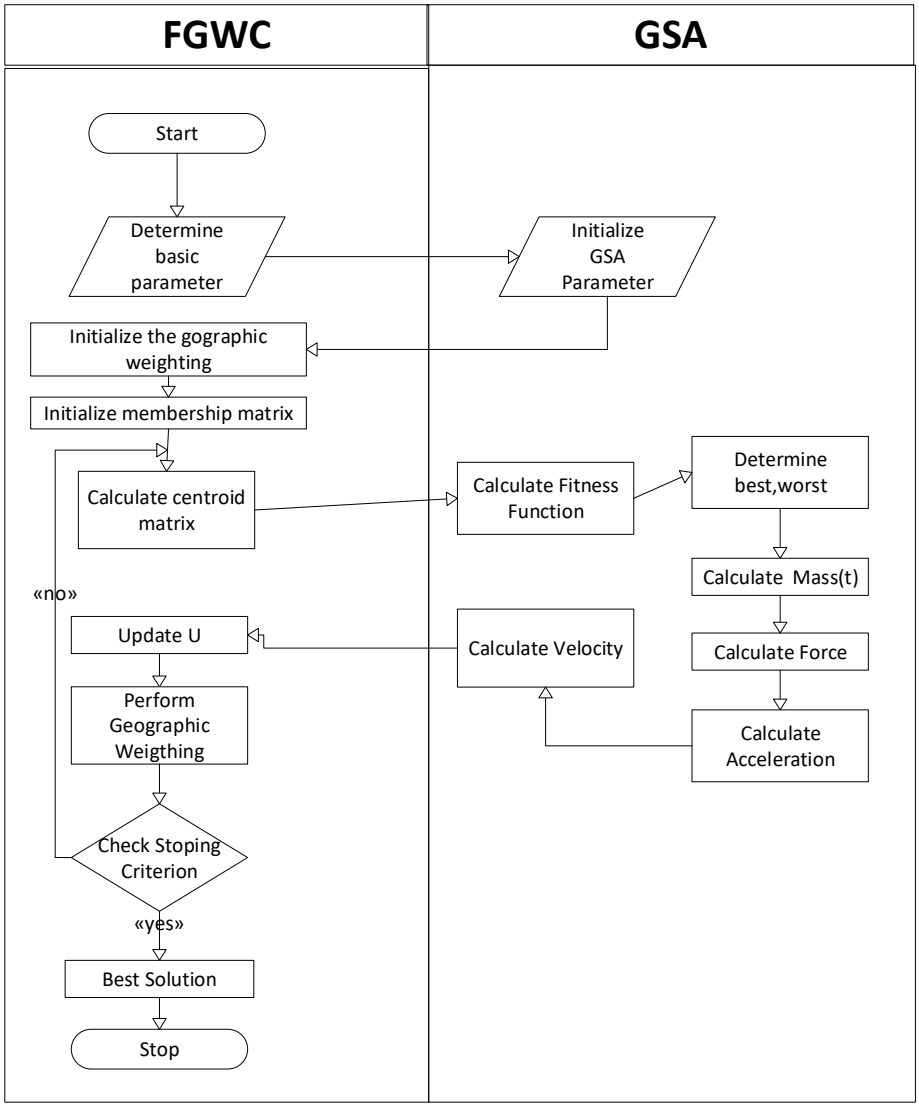

Figure 2 Flow chart of proposed method

Cluster Validity Index

The main problem of cluster validity is finding objective criterion to determine the partition value from the clustering algorithm [7]. In this research we use several cluster validity index. I.e. Partition Coefficient (PC), Classification Entropy (CE), Separation Index (S), Xie Beni Index, and IFV index.

Partition Coefficient measures the average number of relatively degree sharing of each object in membership matrix. The greater value of PC indicate better clustering quality.

$$PC = \frac{1}{N}\sum_{i=1}^{c}\sum_{j=1}^{N}\mu_{ij}^{2}, \qquad (20)$$

where "μ" _ij is membership degree of item j in cluster i.

The Classification Entropy (CE) index is used to define the fuzziness of partition in each cluster:

$$CE = -\frac{1}{N}\sum_{i=1}^{c}\sum_{j=1}^{N}\mu_{ij}\log_a(\mu_{ij}). \qquad (21)$$

The Separation Index and Xie Beni Index measure the compactness and the separation of each cluster. The minimum value of Separation Index and Xie Beni Index indicate the better clustering validity.

TABLE 1.
VARIABLES USED IN THE ANALYSIS

Variables	Description
Literacy rate	Proportion of population of certain age group that can read and write Latin letters, letters Arabic, or Other letters
Mean years of schooling	Mean of population
Expected mean years schooling	Average number of years spent by population aged 15 years for formal education
Net enrolment rate primary school	Proportion of population of primary school age that actually attend primary school
Net enrollment rate junior high school	Proportion of population of junior high school age that actually attend junior high school
Primary school Teacher and student ratio	Proportion of number of teacher and student
Primary school student and school ratio	Proportion of number of teacher and student
Junior high school teacher and student ratio	Proportion of number of teacher and student
Junior high school student and school ratio	Proportion of number of teacher and student
Junior high school dropout rate	Proportion of population of junior high school age that dropped out from school
Average monthly expenditure per capita	Average monthly expenditure per capita especially for education

$$S = \frac{\sum_{i=1}^{c}\sum_{j=1}^{N}\left(\mu_{ij}\right)^{2}\left\|x_j - v_i\right\|^{2}}{N\, min_{i,k}\|v_k - v_i\|^{2}}, \quad (22)$$

$$XB = \frac{\sum_{i=1}^{c}\sum_{j=1}^{N}\left(\mu_{ij}\right)^{m}\left\|x_j - v_i\right\|^{2}}{N\, min_{i,j}\left\|v_j - v_i\right\|^{2}}, \quad (23)$$

where v_i is the centroid of the cluster i.

To validate the cluster fuzziness in spatial data, IFV is implemented as it is stable and robust [8]. Higher IFV index shows better result.

$$IFV = \frac{1}{c}\sum_{j=1}^{c}\left\{\frac{1}{N}\sum_{k=1}^{N}\mu_{kj}^{2}\left[\log_2 c - \frac{1}{N}\sum_{k=1}^{N}\log_2 \mu_{kj}\right]^{2}\right\}. \quad (24)$$

2. Methods

The Improved FGWC using GSA

As mentioned before cluster center initialization in FGWC could fall in local optimal solution easily affecting the clustering results. We propose to minimize the objective function by using GSA to initialize the initial centroid. This is the objective function that will be minimized.

$$J_m(\boldsymbol{U},\boldsymbol{V}) = \sum_{k=1}^{N}\sum_{i=1}^{c}(\mu_{ik})^{m}\|y_k - v_i\|^{2}. \quad (24)$$

Here is step by step of the proposed methods:

Step 1: Determine the basic parameter, number of cluster c, degree of fuzziness m, threshold of error, maximum number of iteration, and some parameter for weighted function.

Step 2: Initialize the GSA parameter such as gravity constant G.

Step 3: Initialize the geographic weighting.

Step 4: Initialize membership matrix.

Step 5: Calculate centroid matrix.

Step 6: Calculate fitness function and start to optimize FGWC using GSA.

Step 7: Update membership matrix

Step 8: Perform geographic weighting to membership matrix

Step 9: Repeat step 5-8 until reach stopping criteria.

For more detail, the step can be seen at Figure 2

We compare the proposed method with the standard FGWC and different optimization approaches such as Particle Swarm Optimization, Artificial Bee Colony and Simulated Annealing, using a case study of Educational Profile of Jawa Tengah Province 2015 published by BPS' Statistics of Jawa Tengah Province, Indonesia. The variables were selected based on research conducted by Bustomi in 2012 [9] which conclude that inequality education in Central Java Province caused by 4 dimensions.

According to the dimension, we choose 11 variables that represent civil participation, education quality and facilities. The details of each variable is presented in Table 1. Hence the dataset used contains 11 variables of 35 regencies in Central Java Province, Indonesia.

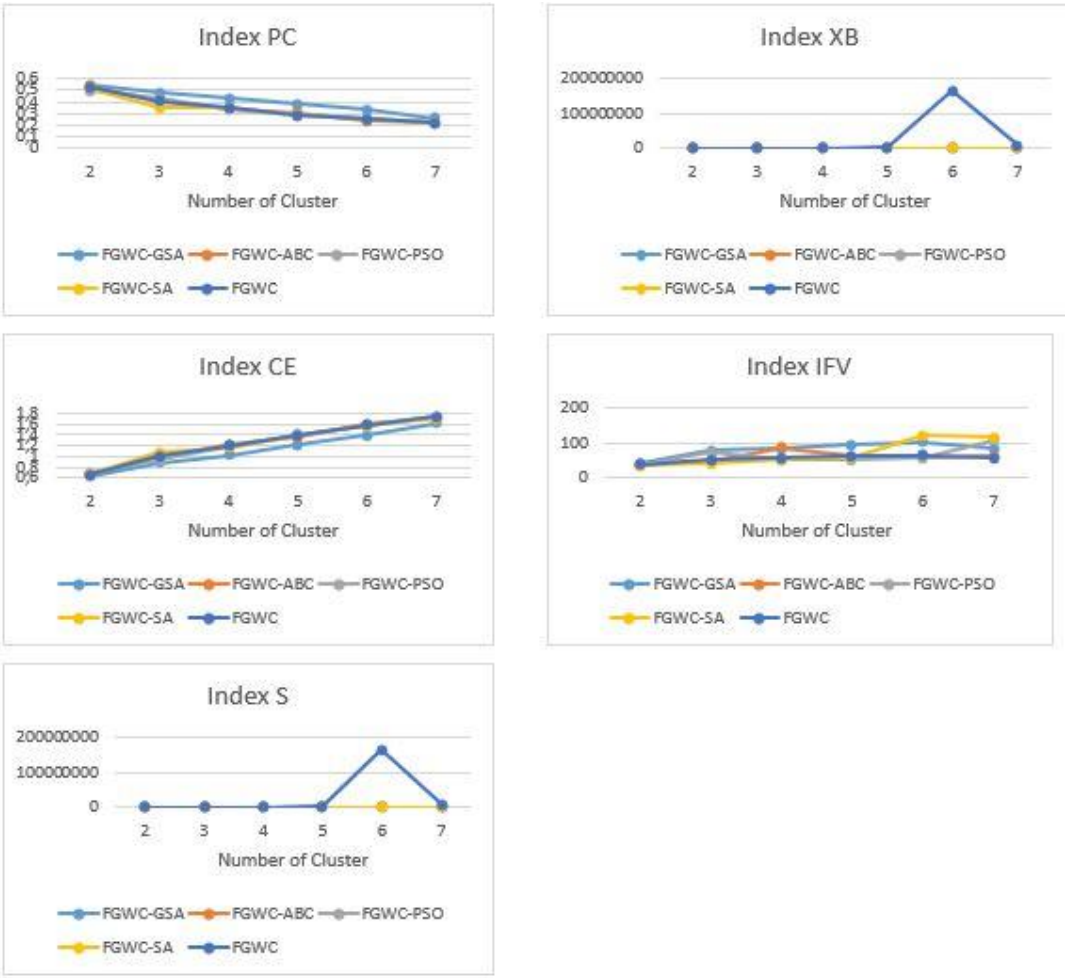

Figure 3. The result of Cluster Validity in different number of clusters obtained from standard FGWC, FGWC-GSA and the other optimization approaches.

The proposed method is implemented in R and now available in CRAN (spatialClust package). The Graphical Interface is also available in FAST [10] which can be accessed through www.stis.ac.id/fast.

3. Results and Analysis

Figure 3 shows the results of different cluster validity indexes from the case study for standard FGWC, FGWC-GSA and other approaches. The x-axis is different number of clusters, and the y-axis is the corresponding validity index.

It can be seen that FGWC-GSA give higher Partition Coefficient Index, IFV index and as Classification Entropy (CE) compared to the other approaches. Furthermore, the FGWC-GSA provide lower Separation Index and Xie Beni Index compared to the others.

In general we observed that FGWC-GSA outperforms FGWC and the other optimization approaches in all validity indexes and all number of clusters.

Based on the results of the analysis, areas in Central Java can be grouped into three clusters based on educational indicators. Cluster 3 contains regencies (such as Semarang and Salatiga) with the high educational quality. While the cluster 1 (e.g., Cilacap, Purbalingga) is the cluster with poor education quality. Cluster 2 consisting medium education quality such as Sragen and Banyumas.

4. Conclusion

In this research, we proposed a new method to avoid local optimal solution that may occur in initial phase of centroid in FGWC, using Gravitational Search Algorithm (GSA) approach. The results show that the proposed method out performs the standard FGWC and its other modification in terms of cluster validity.

References

[1] W. W. Arie, Improvement of Fuzzy Geo-Demographic Clustering using Metaheuristic Optimization on Indonesia Population Cencus, Bandung: ITB, 2014.

[2] A. Mulyanto and R. S. Wahono, "Penerapan Gravitational Search Algorithm untuk Optimasi Klasterisasi Fuzzy C-Means," Jurnal of Intelligent System,vol.1,no.1, pp. 42-47, 2015.

[3] G. A. Mason and R. D. Jacobson, "Fuzzy Geographically Weighted Clustering," in in Proceedings of the 9th International Conference on Geocomputation, 2007.

[4] J. C. R. E. &. W. F. Bezdek, " FCM: The Fuzzy c-meansCLustering Algorithm," Computers & Geosciences, vol. 10, pp. 191-203, 1984.

[5] R. Khadanga and S. Panda, "Gravitational search algorithm for Unified Power Flow Controller based damping controller design," 2011 International Conference on Energy, Automation and Signal, pp. 1-6, 2011.

[6] E. Rashedi, H. Nezamabadi-pour and S. Saryazdi, "GSA: A Gravitational Search Algorithm," Information Sciences, vol. 179, no. 13, p. pp. 2232–224, 2009.

[7] C. Oscar, R. Elid, S. Jose and N. Enrique, "Optimization of the Fuzzy C-Means Algorithm using Evolutionary Methods," Engineering Letters, 20:1, EL_20_1_08, 2012.

[8] L. H. Son, B. C. Cuong, P. L. Lanzi and N. T. Thong, "A novel intuitionistic fuzzy clustering method for geo-demographic analysis2," Expert Syst. Appl., vol. 39, no. 10, p. 9848–9859, 2012.

[9] M. J. Bustomi, "Ketimpangan Pendidikan Antar Kabupaten/Kota dan Implikasinya di Provinsi Jawa Tengah," Economics Development Analysis, 2012.

[10] D.S. M Dalimunthe L. et.al. (2014). FAST: a Web-Based Statistical Analysis Forum. Proceeding Islamic Countries Conference on Statistical Science 13 (pp.185-200)