Journal of Physics: Conference Series

PAPER • OPEN ACCESS

Fuzzy c-means and gath-geva methods in clustering districts based on human development index (hdi) in south sulawesi

To cite this article: S Annas *et al* 2019 *J. Phys.: Conf. Ser.* **1317** 012014

Fuzzy c-means and gath-geva methods in clustering districts based on human development index (hdi) in south sulawesi

S Annas*, S Nyompa, R Arisandi, M Nusrang, and Eka S

Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Negeri Makassar, Indonesia

*suwardi_annas@unm.ac.id

Abstract. District grouping in South Sulawesi based on the Human Development Index (HDI) indicators needs to be done as a material for planning and evaluating the targets of government work programs. This grouping is based on dominant indicators of the high and low HDI. The value of the HDI indicator needs to be considered so that the achievement of each indicator is known. Statistical analysis that can be used to group districts that have similarities is cluster analysis. The method that is currently developing is fuzzy clustering analysis, which classifies objects using certain membership degrees. Fuzzy clustering algorithm that can be used is Fuzzy C-means (FCM). Another method of fuzzy clustering analysis developed further is Gath Geva (GG), which is able to detect groups with different forms. In this study, the fuzzy clustering process on the FCM and GG methods with the same parameters and shows that the GG method is better than the FCM method. This conclusion is based on a total of 1000 iterations. The GG method gives an objective function value smaller than FCM, besides it gives a faster-conferencing iteration result.

1. Introduction

The success of development is not only measured by the high economic growth but also by the improvement of human quality [5.6]. Before the 1970s, the success of development was solely measured by the Gross National Product (GNP) growth rate. In fact, there are still many countries with high GNP growth rates have low human quality. Human development measurement was first introduced by the United Nations Development Program (UNDP) in 1990. UNDP introduced a new idea in the measurement of human development referred to as the Human Development Index (HDI) [7.8.9]. Since then, HDI has been published regularly in the annual Human Development Report (HDR).

The HDI explains how residents can access development results in obtaining income, health, education, etc. [11.15]. The calculation of HDI is based on 4 indicators that have different measurement units, in which the health index is measured by the variable life expectancy birth (LEB), the education index is measured by the variable mean years of schooling (MYS), expected years of schooling (EYS), and gross national income per capita (GNI) measured by variable per capita expenditure [3.5.6].

According to the Central Bureau of Statistics, Indonesia's HDI in 2015 was 68.90% and has been on moderate status. Especially in South Sulawesi which houses 24 second-level regions with 21 regencies and 3 cities [3] had HDI in 2013 which was below the national HDI (68.31), ranked 15th nationally with a score of 67.92. The high and low HDI of the districts/city is only indicated by the composite index, but it is not shown which indicator is dominant towards the high/low HDI rating.

However, the value of each indicator forming the HDI needs to be seen in order to know the achievement of each indicator. Clustering regencies/cities in South Sulawesi based on HDI indicators need to be carried out as material for planning and evaluating the objectives of government programs to improve human development rates. Cluster analysis is a multivariate analysis technique that can be used to group objects in such a way that objects in one group are very similar and objects in various groups are quite different [14.16.19].

Cluster analysis can be divided into two methods, hierarchical and non-hierarchical clustering methods [17.10]. In a hierarchical or non-hierarchical clustering process, group formation is carried out in such a way that each object is right in one group. However, at some point, it cannot be performed to place an object precisely in a group, because actually the object is located between two or more other groups. So it is necessary to do clustering using Fuzzy C-means which takes into account the level of fuzzy set membership as the basis for weighting [4.18]. By using this technique, objects tend to become members of a group where the object has the highest degree of membership towards the group [12.19].

The extension for the Fuzzy C-means model is proposed by Gustafson and Kessel, where the distance between objects with group centers is calculated using the Mahalanobis distance formula [13]. Another fuzzy variant that has been developed is the Gath-Geva clustering model. This algorithm uses the Mahalanobis distance formula but with the addition of the fuzzy covariance matrix [1.2].

2. Material and Method

2.1. *Material*

In this study, we used secondary data published by the Central Bureau of Statistics of South Sulawesi Province, namely the indicator data of HDI. We used 4 variables, those are the average estimates of many years that can be taken by someone from birth (LEB), the average school duration (MYS), the school duration (in years) expectation by children at a certain age in the future (EYS), and the average monthly real household expenditure divided by average number of household members (GNI). The objects of this study are 24 districts/cities in South Sulawesi Province, consisting of 3 cities and 21 districts.

2.2. *Method*

2.2.1. *Determination of Optimum Group Amount*

- FCM clustering and Gath-Geva (GG) algorithm
  a. Determining the number of groups (c)
  b. Determining the initial initiation of the U partition matrix randomly and calculating the centroid or v_i from each group.
  c. Calculating the distance measure which is square Euclidean distance according to the equation

$$\boldsymbol{D}_{ik} = \boldsymbol{D}(\mathbf{x}_k, \mathbf{v}_i) = \sum_{i=1}^{c} \|\mathbf{x}_k - \mathbf{v}_i\|^2 = \sqrt{\sum_{i=1}^{c} (\mathbf{x}_k - \mathbf{v}_i)^2} \tag{1}$$

In FCM clustering

$$\boldsymbol{D}_{ik} = \frac{(2\pi)^{(\pi/2)}\sqrt{\det(F_{wi})}}{\alpha_i} \exp\left(1/2(\mathbf{x}_k - \mathbf{v}_i)^T F_{wi}^{-1}(\mathbf{x}_k - \mathbf{v}_i)\right) \tag{2}$$

In GG

  d. Calculating the value of the membership data function in each group or updating the U partition matrix

$$\boldsymbol{u}_{ik} = \sum_{j=1}^{c} \left[\frac{\boldsymbol{D}(\mathbf{x}_k, \mathbf{v}_i)}{\boldsymbol{D}(\mathbf{x}_k, \mathbf{v}_j)} \right]^{-2/(m-1)} \tag{3}$$

and allocating the data to the nearest centroid and calculating the new centroid.

e. Back to step 2 when the change in the value of the membership function is still above the threshold value (ε) specified by | U^l - U^{l-1} | <ε or when the change in the value of the objective function is calculated based on

$$J(\mathbf{X}:\mathbf{U},\mathbf{V}) = \sum_{i=1}^{c}\sum_{k=1}^{N}(\boldsymbol{u}_{ik})^m \boldsymbol{D}(\mathbf{x}_k,\mathbf{v}_i)^2 \qquad (4)$$

Is above the specified threshold value. The threshold value is a very small value

2.2.2. Selection of the Best Clustering Method

- Evaluation of Clustering Results

Assessment can be performed by comparing the results of the cluster by using the criteria of two standard deviation values, namely the average standard deviation in the group (S_W) and the standard deviation between groups (S_b). The average standard deviation formula in groups:

$$S_w = K^{-1}\sum_{k=1}^{k} S_k \text{ and } S_b = \left[(K-1)^{-1}\sum_{k=1}^{k}\left(\overline{X}_k - \overline{X}\right)^2\right]^2 \qquad (5)$$

The smaller the S_W value and the greater the S_b value, the clustering method algorithm has better performance, so the ratio between S_W and S_b is used. The smallest S_W / S_b ratio shows the best clustering accuracy.

- Cluster Validity Index

3. Result and Discussion

3.1. Determination of Optimum Group Amount

The clustering of the FCM method was carried out on the number of groups of two to six. Other FCM parameters were set the same for each number of groups, namely the rank (w) of 2, the maximum iteration of 1000 times, and ε of 10^{-10}.

For the GG method, the radius was changed until the number of groups is two to six, this difference is due to the distance used in determining the number of groups. So, if the distance (r) value is enlarged, the number of groups will be smaller, and vice versa. Other GG parameters were set the same for each maximum iteration of 1000 times, and ε for 10^{-10}. Validity index obtained for each number of groups was listed in Table 3.1.

Table 3.1. Index Validity of FCM and GG Clustering Results

Method	Number of Groups	Number of Iterations	Objective Function	PC Index
FCM	1	30	12.88	389.283
	2	40	11.13	390.045
	3	30	8.90	289.876
GG	1	18	11.71	367.23
	2	38	10.21	381.70
	3	31	7.20	242.56

Table 3.1 shows the result of the FCM method, that is the optimum Partition Coefficient (PC) index value of 390.045 with the number of groups is 2, the iteration process stopped at the 40rd iteration because the value of | Pt - Pt-1 | <ξ. The objective function value in the last iteration obtained is 11.13. While the PC index in GG method is the value of the objective function during the iteration. In this study, the optimum PC is 381.70 with the number of groups is 2, the iteration process stopped at the 38th iteration because the value of | Pt - Pt-1 | < ξ. The objective function value in the last iteration obtained is 10.21. The number of members of the first group up to the third group in a row are 10, 3, 11.

The group center value for both methods:

In the last iteration (40rd iteration), the center of the V_{kj} group generated with k = 1,2,3,4 and j = 1, 2, 3 are:

Central Value of Indicator Groups Forming HDI with the FCM method

$$V_{kj} = \begin{pmatrix} 69{,}88 & 15{,}03 & 11{,}05 & 14300{,}87 \\ 66{,}01 & 13{,}04 & 6{,}86 & 9505{,}99 \\ 70{,}99 & 11{,}99 & 6{,}93 & 9667{,}80 \end{pmatrix} \quad (6)$$

This value is the coordinates of the three center points of the group and gives an outline of each group, namely:

Group 1 consists of districts / cities with LEB of 69.88 (years), EYS of 15.03 (years), MYS is 11.05 (years), GNI is 14300.87 (rupiahs).

Group 2 consists of districts / cities with LEB of 66.01 (years), EYS is 13.04 (years), MYS is 6.86 (years), GNI is 9505.99 (rupiahs).

Group 3 consists of districts / cities with LEB of 70.99 (years), EYS of 11.99 (years), MYS of 6.93 (years), GNI is 9667.80 (rupiahs).

The central value of the indicator forming group with the GG method

$$V_{kj} = \begin{pmatrix} 70{,}01 & 12{,}02 & 6{,}98 & 8964{,}01 \\ 70{,}04 & 13{,}76 & 11{,}55 & 1409{,}23 \\ 66{,}98 & 13{,}09 & 7{,}77 & 9709{,}98 \end{pmatrix} \quad (7)$$

Group 1 consists of districts / cities with LEB of 70.01 (years), EYS of 12.02 (years), MYS of 6.98 (years), GNI is 8964.01 (rupiahs).

Group 2 consists of districts / cities with LEB that is 70.04 (years), EYS is 13.76 (years), MYS is 11.55 (years), GNI is 1409.23 (rupiahs).

For group center 3 consists of districts / cities with LEB of 69.98 (years), EYS is 13.09 (years), MYS is 7.77 (years), GNI is 9709.98 (rupiahs).

The degree of membership of each district/city is shown in Table 3.2 and it obtained the information about the tendency of districts/cities to enter certain groups. The greatest degree of membership shows the highest tendency of districts/cities to become members of the group.

Table 3.2. Degree of Districts/City Membership for Each Group on the Last Iteration using the FCM and GG Methods

Method	South Sulawesi Region	Membership Degree (μ) For Each Cluster On Last Iteration		
		μ_{i1}	μ_{i2}	μ_{i3}
FCM	Selayar	-1.8651	-1.1735	-1.6314
	Bulukumba	-1.8845	-0.9391	-1.8465
	⋮	⋮	⋮	⋮
GG	Selayar	-0.7284	-0.9621	-0.2705
	Bulukumba	-0.9434	-0.9815	-0.0361
	⋮	⋮	⋮	⋮

The complete results of clustering 24 regencies/cities into 3 groups are shown in Table 3.3. and Figure 3.1.

Table 3.3. Results of Districts/ Cities Clustering in South Sulawesi Province using FCM and GG

Method	Group	Members of the group
	1	Kota Makassar, Kota Parepare, dan Kota Palopo,
	2	Selayar, Bulukumba,Bantaeng, Jeneponto, Takalar, Sinjai, Pangkep, Bone, Soppeng, Wajo, dan Luwu Utara,

FCM	3	Gowa, Maros, Barru, Sidrap, Pinrang, Enrekang, Luwu, Tana Toraja, Luwu Timur, dan Toraja Utara.
GG	1	Gowa, Maros, Barru, Sidrap, Pinrang, Enrekang, Luwu, Tana Toraja, Luwu Timur, dan Toraja Utara,
	2	Kota Makassar, Kota Parepare, dan Kota Palopo,
	3	Selayar, Bulukumba, Bantaeng, Jeneponto, Takalar, Sinjai, Pangkep, Bone, Soppeng, Wajo, dan Luwu Utara.

3.2. Selection of the Best Clustering Method

The comparison of the optimal clustering results of the two methods is listed in Table 3.4.

Table 3.4. Comparison of Optimum FCM and GG Results for South Sulawesi HDI

Method	Number of Group	Number of Iteration	Objective function	PC Index
FCM	2	40	11.13	390.045
GG	2	38	10.21	381.700

Table 3.4 shows that the FCM method gave better clustering results based on the PC index. However, when viewed as a whole with 1000 iterations, the GG method gave an objective function value smaller than FCM, the speed of the GG method gave faster results shown from the number of iterations of the GG method giving less value.

By using the GG method after the group was formed to 24 districts / cities, taken all research objects of the average of the HDI-forming variables, namely LEB, EYS, MYS, and GNI ($\bar{X}$). Furthermore, each group was taken on average for variables LEB, EYS, MYS, and GNI ($\bar{X}$c), in each variable in the group, was marked, if ($\bar{X} \leq \bar{X}$c) then given a positive sign (+), whereas if ($\bar{X} > \bar{X}$c) then given a negative sign (-)

Table 3.5. Group Characteristics Based on Average

Group	LEB	EYS	MYS	GNI
1	+	+	-	-
2	+	+	+	+
3	-	-	-	-

Table 3.5 shows that group 1 variables that have an average value of HDI above the average value in the South Sulawesi Province are the life expectancy and school duration. Therefore, in group 1 the variables that need more attention are the variables of average school length and purchasing power parity. Group 2 has the average value of HDI above the average value in South Sulawesi Province. Group 2 is the best group model. Group 3 variables have an average value of HDI below the average value in South Sulawesi Province, hence group 3 has the lowest HDI. Figure 3.1 shows the results of district / cities grouping in South Sulawesi with the best method

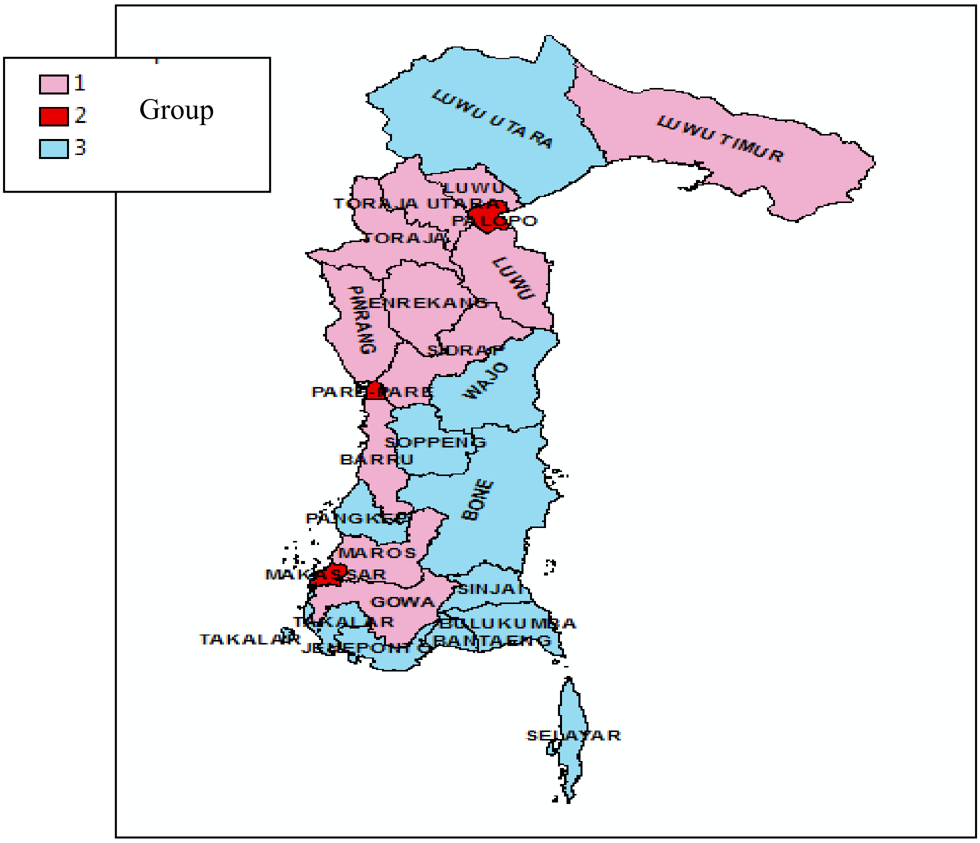

Figure 3.1. Map of South Sulawesi Province Based on Clustering of The Human Development Index

4. Conclusion

The optimal number of groups, based on PC validity index, are 3 groups, both using the FCM and GG methods. However, based on the objective function criteria and the number of iterations, the GG algorithm is better than FCM. In addition, the results of the analysis show that group 2 has an average HDI score above the average value in South Sulawesi Province. While group 3 is the group with the lowest HDI.

Acknowledgement

The authors would like to express their gratitude to the Central Bureau of Statistics of South Sulawesi Province, who have given permission and help the data collection of HDI, so that this research runs smoothly.

References

[1] Abonyi, J. dan Szeifert, F. 2003. *Supervised Fuzzy Clustering for the Identification of Fuzzzy Classifiers*. Journal Elsevier. Vol. 24. pages. 2195-2207.

[2] Abonyi, J. dan Feil, B. 2007. *Cluster Analysis for Data Mining and System Identification*. Birkhäuser. Berin.

[3] BPS. 2015. *Human Development Index 2010-2014*. Jakarta. Badan Pusat Statistik.

[4] Bezdek, J.C. 1981. *Pattern Recognition with Fuzzy Objective Function Algorithms.* Plenum Press. New York.

[5] BPS. 2011. *Human Development Index 2009-2010 Linkages between HDI, IPG, and IDG.* Jakarta. CV.Rioma

[6] Badan Perencanaan Pembangunan Daerah. 2015. *South Sulawesi Human Development Analysis.*

[7] Barro, R, J. 1998. *Human Capital and Growth in Cross Country Regressions. Journal of Economics*. Jurnal of Economics Harvard University. No. 214.

[8] Barro, R. J dan Martin. 1999. *Economic Growth*. MIT Press

[9] Becker, G, S. 2002. *Investment in Human Capital. A Theory Analysis* . The Journal of Political Economy. Vol 70. The University of Chicago Press.

[10] Bacher et al. 2004. *SPSS Two Step Cluster. A First Evaluation*. Nuremberg. University of Erlangen

[11] Commision on The Growth and Development. 2008. *The Growth Report. Strategies for The Sustained Growth and Inclusive Development.* Washington DC. World Bank

[12] Eka. S. 2017. *Subtractive Fuzzy C-Means (SFCM) Method in Grouping of Regencies/Cities in South Sulawesi Province Based on Human Development Index Indicators* Skripsi UNM.

[13] Gustafson, D. and Kessel, W. 1979. *Fuzzy clustering with a fuzzy covariance matrix*. Proceedings of the IEEE CDC. San Diego. CA, USA. pages. 761-766.

[14] Kim, D.W., Lee, K.H., Lee, D. 2003). *Fuzzy cluster validation index based on inter-cluster proximity.* Pattern Recognition Lett., No.24, pages. 2561-2574.

[15] Khodabakhshi, A. 2011. *Relationship between GDP and Human Development Indices in India.* International Journal of Trade. Economics and Finance. Vol. 2, No. 3.

[16] Maxwell, B.A., Pryor F.L., dan Smith C. 2002. *Cluster Analysis In Cross-Cultural Research.* International Journal of World Cultures. Vol. 1. No.13. pages. 22-38.

[17] Mongi, Charles.E. 2015. *Use of Two Step Clustering Analysis for Mixed Data*. JdC Vol.4 No.1.

[18] Pedrycz, W. 2007. *Advances in Fuzzy Clustering and its Applications.* University of Alberta, Canada Systems Research Institute of the Polish Academy of Sciences, Poland. Ltd. ISBN: 978-0-470-02760-8.

[19] Shihab, A. I.2000. *Fuzzy Clustering Algorithm and Their Applicaion to Medical Image Analysis*. University of London. London.