Implementation of the Gath-Geva Clustering Algorithm in the Clustering Districts/Cities in Central Sulawesi Based on Public Health Development Indicators

Nichen Grasya Peso’a(✉), Rais, and Nurul Fiskia Gamayanti

Department of Statistics Faculty of Mathematical and Natural Sciences, Tadulako University,
Jl. Soekarno Hatta KM.9, Palu 94118, Indonesia
nichenpesoa26@gmail.com

Abstract. Public health development indicators are a health indicator set that can describe health problems. The health indicators set can, directly and indirectly, increase longevity and healthy life expectancy. Central Sulawesi was listed as the sixth province in Indonesia with the lowest public health development score. In order to develop an equitable and appropriate health development strategy in each region in Central Sulawesi, this can be done by clustering the districts/cities in central Sulawesi on the basis of public health development indicators using the Gath-Geva Clustering method. The algorithm of the Gath-Geva Clustering method uses a distance norm based on Fuzzy Maximum Likelihood Estimation (FMLE) with an exponential aspect, which allows the method to converge faster and thus reduce the number of iterations. Using the validity index Kwon, rank/weight $m = 2$, threshold $\varepsilon = 0.0005$, and the iterations maximum number is 1000, we obtained that the optimal number of clusters is 3 clusters, and the attributes of each cluster vary according to the development index of public health. Cluster 1 is a low public health development cluster, cluster 2 is a medium public health development cluster, and Cluster3 is a high public health development cluster.

Keywords: Clustering · Gath-Geva Clustering · Public Health Development · Kwon

1 Introduction

Healthy development is an effort that a country must make to prevent sudden death and also increase the awareness, willingness, and ability of society to live a healthy life thus achieving public health at a high level. The level of health itself is one of the investments that have an important role in developing human resources productivity in the society and the economy, and any situation that causes health problems in Indonesian society will bring huge economic losses to the country. The collection of a health indicators series called the public health development index, can describe health problems through simple and direct measurements. The selected public health development indicators can be recommended as a reference for a health development plan in Indonesia [1].

© The Author(s) 2023
N. I. Inda et al. (Eds.): ISST 2022, APR 7, pp. 320–328, 2023.
https://doi.org/10.2991/978-94-6463-228-6_36

Referring to the 2018 Public Health Development Index Handbook [1], Central Sulawesi was listed as the sixth province in Indonesia with the lowest score in the Public Health Development Index. This shows that health development in Central Sulawesi still cannot be said to be good. Equitable health development in each region in Central Sulawesi Province needs to be done to increase the value of Public health development index in Central Sulawesi. To assist the government in making decisions and strategizing health development that is equitable and targeted at each region in Central Sulawesi Province, clustering can be done based on the characteristics of health development.

One method for clustering is the Gath-Geva Clustering method, where the algorithm of this method involves exponential aspects in the distance formula which makes this method have a faster convergence rate so that the number of iterations becomes less [2].

Based on the explanation discussed above we can see the urgency to fix the problem of public health development in Central Sulawesi, so the author is interested in applying the Gath-Geva Clustering Algorithm in Clustering Districts/Cities in Central Sulawesi based on public health development indicators.

2 Materials and Method

2.1 Clustering

Clustering is a method for classifying data and performing a segmentation of data where the samples are grouped into clusters and every object in a cluster is more likely to be related or have similar characteristics than objects in different clusters [3]. The main purpose of clustering methods is to group a set of data or objects into clusters so that each cluster can contain as similar data as possible [4].

2.2 Determining the Optimal Number of Clusters

The optimal number of clusters can be determined using validity metric. One of the validity measures that can be used is the Kwon validity measure, which is a further development of the Shabeny validity measure. Kwon's effectiveness index is very effective and can overcome the weakness of Shebeni's effectiveness index which decreases monotonically with the increase of the number of clusters and the approach to the dataset [5].

The kwon validity indicator value can be calculated using the following Eq. [5]:

$$GI_K = \frac{\sum_{k=1}^{n}\sum_{i=1}^{c}\left(\mu'_{ik}\right)^{\eta}\left\|x_{ij}-v_{kj}\right\|^2 + \frac{1}{c}\sum_{i=1}^{c}\left\|v_{kj}-\bar{v}\right\|^2}{min\left\|v_k - v_j\right\|^2}$$

which:

GI_K: Kwon index

c: Number of clusters

η: The rank of the weights

μ'_{ik}: Normalized degree of conformity

v_{kj}: Centeroid

$\bar{v}$: Average of centeroid

$v_k - v_j$: Distance between centeroid

The optimum cluster criterion is given by the minimum kwon value.

2.3 Gath-Geva Clustering

The Gath-Geva Clustering (GG) algorithm is a further development of Fuzzy C-Means (FCM) and Gustafson-Kessel (GK), where Gath and Geva [2] further observed that the algorithm is able to cluster different shapes to detect size and data density. The algorithm uses a distance norm based on Fuzzy Maximum Likelihood Estimation (FMLE) with an exponential aspect, which allows the method to converge faster and thus reduce the number of iterations [6].

Here is the algorithm of Gath-Geva Clustering [6]:

1. Enter the data x_ij into an $n \times l$ matrix, where n is the number of observations to group and l is the number of variables.
2. Perform parameter initialization by determining the desired number of cluster simulations (c) with the criteria $2 \leq c < n$. Determining the rank/weight ($m > 1$), a good rank/weight used in this method is $m = 2$ [7]. Determine the maximum iteration (t_{max}) and determine the threshold value (ε).
3. Randomly determine the initial partition matrix U (u_{ik}). u_{ik} is the degree of membership, which refers to how likely the data belongs to a certain group, where $i = 1, 2, \ldots n$ is the number of observations, and $k = 1, 2, \ldots c$ is the number of clusters. Here is the initial form of the partition matrix:

$$u_{ik} = \begin{bmatrix} u_{11} & \ldots & u_{1c} \\ \vdots & \ddots & \vdots \\ u_{n1} & \ldots & u_{nc} \end{bmatrix} \quad (2)$$

$$\sum_{k=1}^{c} (u_{ik}) = 1$$

4. Calculate the centeroid (v_{kj}) with the following equation:

$$v_{kj} = \frac{\sum_{i=1}^{n} (u_{ik})^m x_{ij}}{\sum_{i=1}^{n} (u_{ik})^m}$$

which:
u_{ik}: Membership degree of i-th data in k-th cluster
m: Rank/weight
x_{ij}: The i-th data on the j-th variable

5. Calculate the distance measure between the data and the centeroid with the following equation:

$$D_{ik} = \frac{2\pi^{\left(\frac{\pi}{2}\right)} \sqrt{det(F_{wi})}}{a_i} exp\left(\frac{1}{2(x_{ij}-v_{kj})^T F_{wi}^{-1} (x_{ij}-v_{kj})}\right)$$

with (F_{wi}) calculated by the following formula:

$$F_{wi} = \frac{\sum_{i=1}^{n} u_{ik}^m (x_{ij}-v_{kj})^T (x_{ij}-v_{kj})}{\sum_{i=1}^{n} u_{ik}^m}$$

And (a_i) calculated by the following formula:

$$a_i = \frac{\sum_{i=1}^{n} u_{ik}^m}{n}$$

where:
F_{wi}: Fuzzy covariance matrix
a_i: Probability prior
u_{ik}: Membership degree of i-th data in k-th cluster
m: Rank/weight
x_{ij}: The i-th data on the j-th variable

Table 1. Kwon Validity Index Value

Number of Clusters	Index Validity Kwon
2	242.878
3	**3.682**
4	7.698
5	10.327

v_{kj}: Centeroid
6. Calculating the objective function

$$J_{GG} = \sum_{i=1}^{n} \sum_{k=1}^{c} \left(u_{ik}^{m}\right) D_{ik}$$

which:
u_{ik}: Membership degree of i-th data in k-th cluster
m: Rank/weight
D_{ik}: Distance between the data and centeroid
7. Calculating the new membership degree value:

$$u_{ik} = \left[\frac{(D_{ik})^{\frac{1}{(m-1)}}}{\sum_{k}^{c} (D_{ik})^{\frac{1}{(m-1)}}} \right]^{-1}$$

where:
D_{ik}: Distance between the data and centeroid
m: Rank/weight
8. Return to step 4, if the change in data membership function value is still above the threshold value (ε). The threshold value is a very small value close to zero, the smaller the better but generally used is $\varepsilon = 10^{-5}$.
With the following stopping criteria:
If $|J_t - J_{t-1}| \leq \varepsilon$ or $> t_{max}$, then stop
If not, then the iteration (t) increased $t = t + 1$, repeat step 4.

3 Result and Discussion

3.1 Determining the Optimum Number of Clusters

The optimal number of clusters in this study was determined according to the criteria given by the value of the Kwon Validity Index. Below are the results obtained.

The optimal number of clusters is represented by the minimum weight validity index value, so it can be seen from Table 1 that the minimum weight validity value is in cluster 3, with a value of 3.682. So in this paper 3 clusters are used.

3.2 Determining Parameter Initialization

Firstly, the Gath-Geva Clustering algorithm determine the initialization of the used parameters. The parameter initialization in this study is 3 clusters that are obtained from the results of the kwon validity index value shown in Table 1. The rank/weight (m) used in this study is $m = 2$, the maximum iteration is 1000, the stopping criteria or threshold value used is 0.00001.

Table 2. Initial Membership Degree Value

Data	u_{i1}	u_{i2}	u_{i3}	$\sum u_{ik}$
1	0.215	0.615	0.169	1
2	0.173	0.096	0.731	1
3	0.136	0.034	0.831	1
4	0.350	0.366	0.285	1
5	0.462	0.456	0.082	1
6	0.459	0.508	0.033	1
7	0.062	0.323	0.615	1
8	0.406	0.271	0.324	1
9	0.347	0.330	0.323	1
10	0.042	0.559	0.398	1
11	0.233	0.488	0.279	1
12	0.023	0.664	0.313	1
13	0.490	0.353	0.157	1

3.3 Determining the Initial Membership Degree

The thing to do to determine the initial membership degree value is to generate random numbers u_{ik} with $i = 1, 2, \ldots 13$ (number of data) and $k = 1, 2, 3$ (number of cluster), Assume that the number of data items in each row is 1. The generated membership degree values are as follows (Table 2).

3.4 Calculating the Centeroid Cluster

Determination of the centeroid cluster aims to determine the distance of a data to the cluster center, a data object is included in a cluster if it has the closest distance to the centeroid cluster. The centeroid cluster values obtained are as follows (Table 3).

3.5 Calculate the Distance

The distance measure in the cluster is used to determine the similarity degree of characteristics between objects. This is in accordance with the purpose of the clustering itself, which is to group those that have similarities. The distance measure obtained is as follows (Table 4).

3.6 Calculating the Objective Function

Calculating the objective function aims to find out at what iteration the U and T matrices converge and the iteration is stopped. After the analysis is carried out, the process is declared to stop at the 61st iteration with the objective function obtained of 1.706×10^{-12}.

Table 3. Centeroid Cluster Value

Variables	Cluster 1	Cluster 2	Cluster 3
X1	−0.525	0.952	−0.060
X2	−0.527	0.536	0.298
X3	−0.505	0.550	0.157
X4	−0.313	0.186	0.127
X5	−0.336	0.779	−0.175
X6	−0.445	0.828	−0.249
X7	−0.269	−0.004	0.406
X8	−0.429	1.038	−0.105
X9	−0.293	−0.312	0.880

Table 4. The Distance Measure

Data	*Cluster 1*	*Cluster 2*	*Cluster 3*
1	4.94×10^{-3}	3.10×10^{-13}	3.06×10^{-4}
2	5.89×10^{-5}	2.85×10^{-6}	2.20×10^{-15}
3	7.95×10^{-3}	3.13×10^{-13}	6.69×10^{-10}
4	1.91×10^{-13}	8.89×10^{-6}	8.30×10^{-5}
5	3.70×10^{-13}	7.08×10^{-11}	1.14×10^{-13}
6	1.91×10^{-13}	2.73×10^{-5}	4.75×10^{-5}
7	7.05×10^{-6}	4.04×10^{-12}	6.30×10^{-15}
8	1.06×10^{-5}	2.10×10^{-9}	2.29×10^{-15}
9	1.91×10^{-13}	3.66×10^{-6}	9.30×10^{-6}
10	2.28×10^{-6}	2.72×10^{-6}	2.20×10^{-5}
11	1.91×10^{-13}	8.49×10^{-5}	1.21×10^{-3}
12	1.91×10^{-13}	1.53×10^{-4}	2.11×10^{-5}
13	1.25×10^{-2}	8.81×10^{-13}	3.01×10^{-14}

3.7 Calculating the Membership Degree Value

The calculation of membership values aims to determine the propensity of data to fall into a cluster. The results obtained are shown in the Table 5 below.

From Table 5, it can be seen that each county/city in Central Sulawesi has the membership level value to be a member of the cluster. The membership degree of the data is determined according to the largest membership degree value. The maximum membership value in the table is represented by bold numbers.

Table 5. Membership Degree Value

Data	***Cluster 1***	***Cluster 2***	***Cluster 3***	***Cluster***
1	6.28×10^{-11}	**1**	1.01×10^{-9}	2
2	3.73×10^{-11}	7.71×10^{-10}	**1**	3
3	3.93×10^{-11}	**1**	4.67×10^{-4}	2
4	**1**	2.15×10^{-8}	2.30×10^{-9}	1
5	2.35×10^{-1}	1.23×10^{-3}	$\mathbf{7.64 \times 10^{-1}}$	3
6	**1**	7.00×10^{-9}	4.02×10^{-9}	1
7	8.92×10^{-10}	1.56×10^{-3}	$\mathbf{9.98 \times 10^{-1}}$	3
8	2.17×10^{-10}	1.09×10^{-6}	**1**	3
9	**1**	5.22×10^{-8}	2.05×10^{-8}	1
10	9.64×10^{-10}	8.10×10^{-10}	**1**	3
11	**1**	2.25×10^{-9}	1.58×10^{-10}	1
12	**1**	1.25×10^{-9}	9.05×10^{-9}	1
13	2.33×10^{-12}	3.30×10^{-2}	$\mathbf{9.67 \times 10^{-1}}$	3

3.8 Cluster Averages

The following is the average value of variables in each cluster to determine the characteristics of each cluster.

Based on Table 6, the highest average of a cluster is marked in blue while the lowest average is marked in yellow. The properties of each formed cluster are explained as follows.

1. Cluster 1 consists of 5 districts/municipalities, namely Poso, Toli-Toli, Tojo Una-Una, Banggai Laut, and North Morowali. The districts/cities included in this cluster are

Table 6. Cluster Averages

Variable	***Cluster 1***	***Cluster 2***	***Cluster 3***
X_1	76.180	98.500	88.350
X_2	69.040	83.550	83.883
X_3	69.400	83.850	84.033
X_4	63.904	68.880	72.332
X_5	69.804	79.960	74.028
X_6	76.120	86.950	83.200
X_7	33.580	36.250	37.733
X_8	0.436	4.235	1.257
X_9	245.400	275.500	333.667

areas with the lowest average of all Public Health Development Indicator variables compared to other clusters. This makes the districts/cities in this cluster areas with relatively low public health development.

2. Cluster 2 consists of 2 districts/cities, namely Banggai Islands and Morowali. Districts/cities included in this cluster are areas with an average percentage of Neonatal Visits (KN1), percentage of households that have access to proper sanitation, percentage of households that have used latrines, and percentage of service achievements for hypertension that are high compared to other clusters. This means that districts/cities in this cluster are areas with moderate public health development status.
3. Cluster 3 consists of 6 districts/cities namely Banggai, Donggala, Buol, Parigi Moutong, Sigi, and Palu. The districts/cities included in this cluster are areas with the highest average Pregnancy Check-ups (K4), Percentage of deliveries by health workers, Percentage of population who have at least 1 type of health insurance, Percentage of diarrhea in toddlers treated by health workers, and Number of Midwives compared to other clusters. This cluster has the highest average of the most variables compared to other clusters, so it can be said that this cluster is a cluster with relatively high public health development.

3.9 Clustering Visualization Using Maps

The mapping results of public health development in Central Sulawesi are as follows.

In Fig. 1, you can see the map of Central Sulawesi from the results of the Gath-Geva Clustering analysis where on the map members of cluster 1 are colored red, members

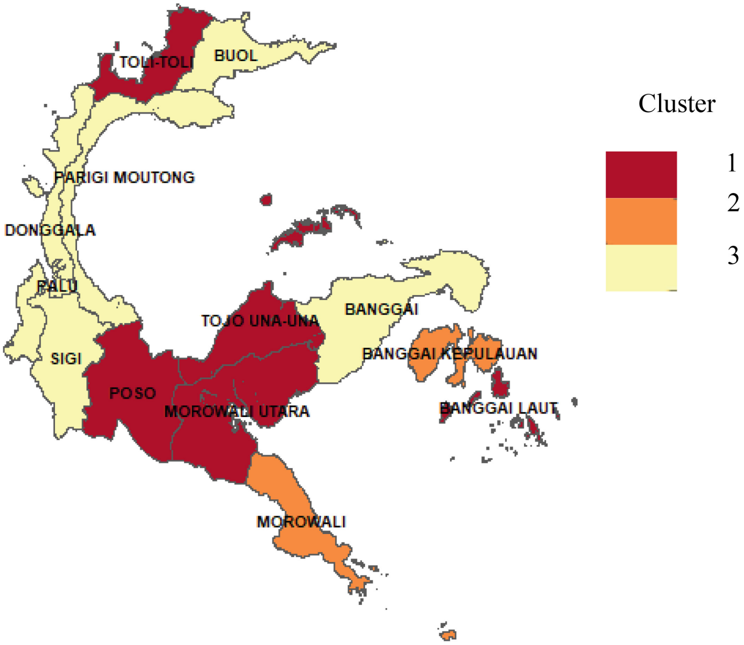

Fig. 1. Mapping of Clustering Results

of cluster 2 are colored orange, and members of cluster 3 are colored yellow. This can make it easier for readers to know the members of the results of the Gath-Geva Clustering analysis that has been done.

4 Conclusion

Based on the results and discussions that have been carried out in this study, it can be concluded that there are 3 clusters with different characteristics of each cluster based on the Public Health Development Indicator. Cluster 1 is a cluster with low public health development, this cluster consists of 5 districts namely Poso, Toli-Toli, Tojo Una-Una, Banggai Laut, and North Morowali. Cluster 2 is a cluster with moderate public health development, this cluster consists of 2 districts, namely Banggai Islands and Morowali. Cluster 3 is a cluster with high public health development, this cluster consists of 6 districts/cities namely Banggai, Donggala, Buol, Parigi Moutong, Sigi, and Palu.

References

1. KEMENKES, "Indeks Pembangunan Kesehatan Masyarakat 2018", Lembaga Penerbit Badan Penelitian dan Pengembangan Kesehatan (LPB), Jakarta (2019).
2. Gath, I., Geva, A.B.: Unsupervised Optimal Fuzzy Clustering, IEEE Trans. Pattern Anal. Mach. Intell., 11, 773–780 (1989).
3. Vialetto, G., Noro, M.: An innovative approach to design cogeneration systems based on big data analysis and use of clustering methods, Energy Convers. Manag., 214, 112901 (2020).
4. Sulastri, H., Gufroni, A.I.: Penerapan Data Mining Dalam Pengelompokan Penderita Thalassaemia, J. Nas. Teknol. Dan Sist. Inf., 3, 299–305 (2017).
5. Kwon, S.H.: Cluster validity index for fuzzy clustering, IEEE Trans. Electron Devices, 44, 1169–1171 (1998).
6. Syoer, R.R., Mashuri, M.: Analisis Kelompok Dengan Algoritma Fuzzy C-Means dan Gath-Geva Clustering, IndoMS J. Stat., 2, 11–26 (2014).
7. Balasko, B., Abonyi, J., Feil, B.: Fuzzy Clustering and Data Analysis Toolbox, Appl. Math. Sci., 183, 49–120 (2005).

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.