ILKOM Jurnal Ilmiah Vol. 14, No. 3, December 2022, pp. 275-283
Accredited 2nd by RISTEKBRIN No. 200/M/KPT/2020; E-ISSN 2548-7779 | P-ISSN 2087-1716

Research Article Open Access (CC–BY-SA)

Hierarchical clustering for crime rate mapping in Indonesia

Rendra Gustriansyah [a,1,*]; Juhaini Alie [a,2]; Nazori Suhandi [a,3]
[a] *Universitas Indo Global Mandiri, Jl. Jenderal Sudirman No. 629, Palembang and 30129, Indonesia*
[1] *rendra@uigm.ac.id;* [2] *juhaini@uigm.ac.id;* [3] *nazori@uigm.ac.id*
** Corresponding author*

Article history: Received March 22, 2022; Revised April 07, 2022; Accepted November 29, 2022; Available online December 20, 2022

Abstract

The Sustainable Development Goals (SDGs) are a blueprint for improving the human life quality. Goal 16 (G16) is related to security, and it is in line with the Universal Declaration of Human Rights and the Preamble to the 1945 Constitution. To support the implementation of the G16 achievement, the Indonesian National Police (Polri) has made serious efforts to provide a sense of safety for the community and to minimize crime rates. One of the efforts that could be made is to map areas based on the level of crimes so that the Polri can determine the appropriate strategy/priority of action for mitigation. Therefore, this study aimed to cluster provinces in Indonesia based on the four G16 indicators of the SDGs related to security, namely the number of homicide cases, the victim proportion, the proportion of people who feel safe walking alone in the area where they live, and the proportion of victims of violence that reported to the police in the past year using five hierarchical clustering methods, namely: Single-Linkage, Average-Linkage, Complete-Linkage, Ward, and Division Analysis. Then, methods were validated and compared using six cluster validations to obtain the most compact method. The results showed that Ward's method outperformed the others and produced three clusters. Clusters 1, 2, and 3 contained 18, 5, and 11 provinces respectively.

Keywords: Crime; Hierarchical Clustering Method; Machine Learning; Sustainable Development Goals

Introduction

The SDGs are a blueprint agreed by all countries in the United Nations (UN), including Indonesia. The agreement is mainly to improve the human life quality worldwide. The SDGs contain 17 Goals and 169 targets to be achieved by 2030. Therefore, the government of Indonesia has committed to realizing the achievement of the SDGs by issuing presidential regulation (Perpres) number 59 of 2017 regarding to the implementation of the SDGs achievement.

Goal 16 (G16) is one of the SDGs related to security aiming at reducing violence and death rates globally. G16 is in line with the universal declaration of human rights stating that everyone has the right for life and freedom and safety. Furthermore, the preamble to the 1945 Constitution states that the government and the country of Indonesia are obliged to protect the entire Indonesian nation and to provide a sense of security to all Indonesian people. Therefore, the Indonesian National Police (Polri) as a state instrument makes serious efforts to provide a sense of safety for the community and minimize crimes based on article 30 section 4 of the second amendment of the 1945 Constitution. One of these efforts is to cluster areas based on the criminal level so that the Polri can determine strategies/priority of appropriate actions for mitigation.

Many studies have clustered provinces/cities based on crime rates. The clustering methods used are kmeans [1]–[3], kmeans and agglomerative clustering [4], [5], kmeans and association rules [6], kmeans and kmedoids [7], kmeans and fuzzy cmeans [8], [9], kmeans and expectation-maximization [10], kmedoids [11], [12], and HDBSCAN [13]. Generally, these studies focus on the k-Means clustering method and do not consider indicators related to the SDGs. This study initiates to fill the research gap by using other clustering methods. In addition, research [5] has inspired the use of various hierarchical clustering methods that have the smallest standard deviation ratio.

For this reason, this study aimed to cluster provinces based on four indicators from the G16 of the SDGs, namely the number of homicide cases (murder), the proportion of the population who are criminal victims (victim), the proportion of the population who feel safe walking alone in their living environment (safe), and the proportion of victims of violence who reported to the police (report). The data is last year data using five hierarchical clustering

http://dx.doi.org/10.33096/ilkom.v14i3.1135.275-283

methods: Single-Linkage (SL), Average-Linkage (AL), Complete-Linkage (CL), Ward (WM), and Divisive Analysis (DA). The selection of these four indicators is based on the availability of data in BPS. While the clustering methods selection to represents the two main categories of the hierarchical clustering method. Furthermore, the clustering methods results are compared and validated using two internal validations (IV): Dunn's index (Di) and silhouette width (Sw), as well as four stability validations (SV), namely average proportion of non-overlap (APN), the average distance (AD), average distance between means (ADB), and figure of merit (FM). This cluster validation aims to identify the most compact and consistent clustering method. Clustering and validation using R programming.

This research can contribute as a reference for the government (Polri) to determine the appropriate strategy/priority of action to mitigate and support the SDGs. Furthermore, this paper is organized in four sections. Section 1 is introduction to the research. Section 2 discusses the research methodology. Section 3 describes the results/findings of the research and discussion. Finally, section 4 provides conclusions and potential future work.

Method

A. Dataset

This study used the 2020 dataset from BPS. **Figure 1** shows the curves of 'murder', 'victim', 'safe', and 'report' in the last year (2020) for each province in Indonesia [14]. In general, this curve depicts security conditions and crime rates in Indonesia based on four indicators from the G16.

The highest number of 'murders' was 99 people in North Sumatra Province, and the lowest was one person in North Maluku Province. The highest proportion of 'victims' was 1.49 in West Nusa Tenggara Province, and the lowest was 0.23 in Bali Province. Furthermore, the highest proportion of 'safe' was in Bali at 81.90 and the lowest in DKI Jakarta at 40.17. Then, the highest proportion of 'report' occurred in West Papua Province at 37.22, and the lowest was in West Nusa Tenggara Province at 12.40.

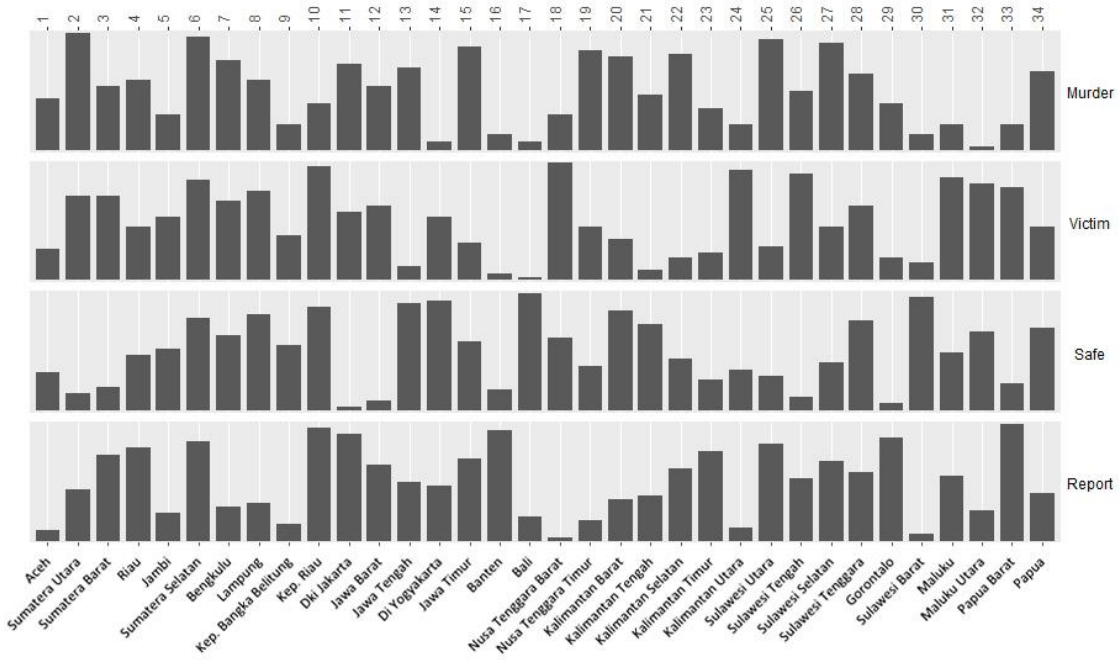

Figure 1. The crime curve in Indonesia is based on four indicators of G16 in 2020 [14].

B. The Proposed Method

In general, the hierarchical clustering method was divided into two categories: the agglomerative (SL, AL, CL, and WM) combining small clusters into a large one and the divisive method (DA) dividing a large one into smaller clusters. The agglomerative started from many clusters where each cluster contained one object. Otherwise, the divisive method started with a large one containing multiple objects.

- Single-Linkage (SL)

  SL is an agglomerative hierarchical clustering method that combines two objects based on the shortest distance of the two clusters sequentially as in (1) to form a complete dendrogram [15].

$$D(c_m, c_n) = min_{ij}\ d(x_i, x_j) \qquad (1)$$

Where $D(c_m, c_n)$ is the distance between c_m and c_n clusters, m and n are the number of objects of cm and cn respectively, x_i an x_j represent the objects in c_m and c_n respectively, and $d(x_i, x_j)$ represents the distance between objects x_i and x_j.

- Average-Linkage (AL)

  AL is an agglomerative hierarchical clustering method that combines two objects based on the average distance of all objects from the two clusters sequentially as in (2) to form a complete dendrogram [15]. Clustering starts from the two objects that are closest to the average distance.

$$D(c_m, c_n) = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} d(x_i, x_j) \quad (2)$$

- Complete-Linkage (CL)

  The opposite of SL, CL is an agglomerative hierarchical clustering method that combines two objects based on the furthest distance from the two clusters sequentially as in (3) to form a complete dendrogram [15].

$$D(c_m, c_n) = max_{ij}\ d(x_i, x_j) \quad (3)$$

- Ward's Method (WM)

  WM is an agglomerative hierarchical clustering method that combines two objects based on the within the sum of square (WSS) value of the two clusters as in (4). The pair of objects that produce the smallest WSS value will be combined first sequentially to form a complete dendrogram [16].

$$D(c_m, c_n) = \frac{mn}{m+n}\|s(x_m) - s(x_n)\|^2 \quad (4)$$

  Where $s(x_m)$ and $s(x_n)$ are the centroids of c_m and c_n, respectively.

- Divisive Analysis (DA)

  DA represents a divisive hierarchical clustering method, the inverse of agglomerative hierarchical clustering. DA starts from one large cluster containing all objects and sequentially breaks down the clusters until each cluster contains only one object as in (5) [17].

$$D(c) = max_{i,j\in C}\ d(x_i, x_j) \quad (5)$$

  Where $D(c)$ is the largest cluster c.

C. Best Number of Clusters (BNCs) Estimation

Estimating BNCs from a dataset can simplify their clustering and validation. For this reason, this study used thirty-two cluster validity indexes (CVI) available in the R-programing package, namely ball, ratkowsky, beale, pseudot2, scott, ccc, hartigan, elbow, ch, kl, tracew, marriot, trcovw, db, duda, cindex, rubin, silhouette, friedman, sdbw, sdindex, dindex, hubert, gap, frey, dunn, ptbiserial, mixture, tau, gamma, gplus, dan mcclain [18]. In addition, Euclidean distance was applied as a distance metric for the dataset with the number of clusters (k) interval from 2 to 10 [19].

D. Cluster Validation

1) Internal Validation (IV)

The IV is a way to measure the quality of clustering results using internal information from the data. The IV value reflects the compactness, relationship and separation of clusters. Compactness is measured by evaluating cluster homogeneity through intra-cluster variance. Relationships indicate the placement of objects in a cluster. Meanwhile, cluster separation is measured by the inter-cluster distance (centroid). The IV includes *Sw* and *Di*.

- The silhouette width

  It measures how compact the cluster is formed based on the average intra-cluster distance and the closest cluster distance [20]. The compact clusters have the *Sw* value close to one. If *Sw* is negative, the object may be located in the wrong cluster.

- Dunn index

  It measures how compact the cluster is formed based on the ratio of the minimum distance between objects that are not in the same cluster using the maximum intra-cluster distance [21]. The compact clusters have a *Di* value close to one. Otherwise, less compact ones have a *Di* value close to zero.

2) Stability Validation

The SV is a way to measure the cluster's sensitivity that is formed. The SV is done by comparing the clustering results using complete data against the clustering results after each column of indicators has been deleted, one by one at a time. SV includes APN, AD, ADB, and FM [22].

- The average proportion of non-overlap (APN) is the average proportion of objects in the same cluster for clustering using complete indicators and without one of them.
- The average distance (AD) is the average intra-cluster distance for clustering using complete indicators and without one of them.
- The average distance between means (ADB) is the average inter-cluster distance (centroid) for clustering using complete indicators and without one of them.
- The figure of merit (FM) is the average intra-cluster variances for the deleted indicator.

The lower the value of APN, AD, ADB, and FM, the lower the sensitivity of the formed clusters.

Results and Discussion

Table 1 presents a descriptive statistical analysis for 34 provinces in Indonesia based on four indicators of the G16 in 2020. The standard deviation indicates that the data is less varied.

Table 1. Descriptive Statistics for The Four Indicators of The G16 in Indonesia in 2020.

Indicator	Observation	Minimum	Maximum	Mean	Std. deviation
Murder	34	1	99	26.41	23.41
Victim	34	0.23	1.49	0.83	0.28
Safe	34	40.17	81.90	64.58	8.55
Report	34	12.40	37.22	22.34	6.41

A. BNCs Estimation

To optimize the estimations of BNCs, outliers were not included in the calculation. **Figure 2** demonstrates a boxplot chart for each indicator with identified outliers. There was one outlier in the 'safe' dataset that was DKI Jakarta, and there were four outliers in the 'murder' dataset, they were North Sumatra, South Sumatra, North Sulawesi, and South Sulawesi.

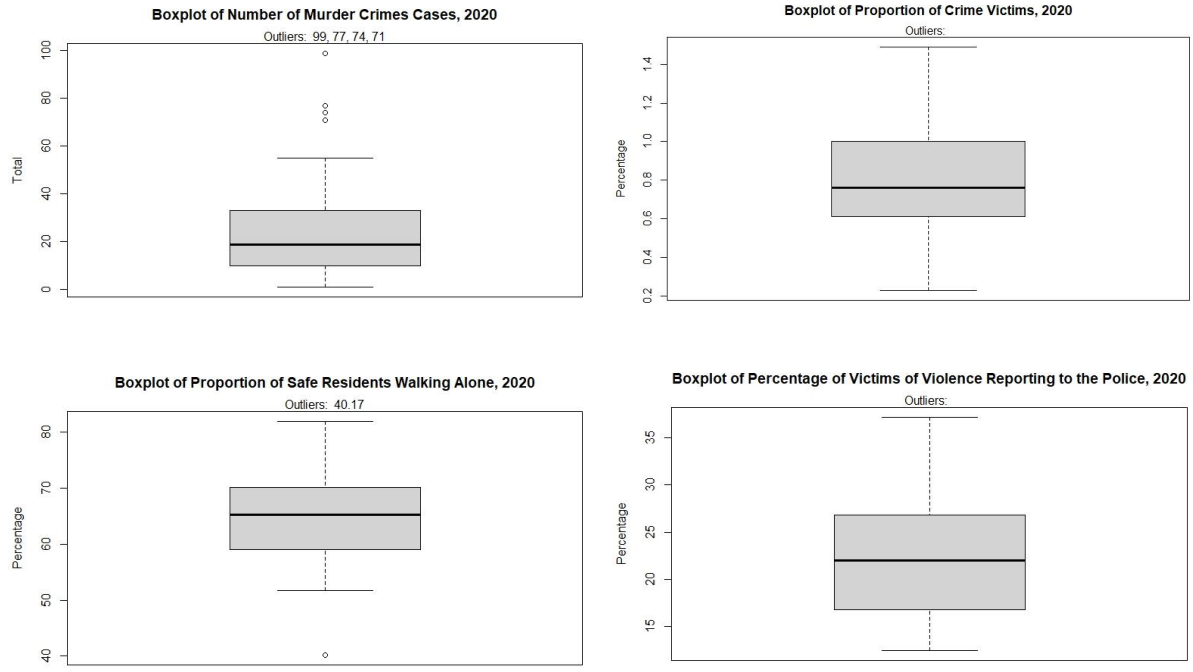

Figure 2. The Boxplot for Each Indicator.

BNCs estimation summarized from thirty-two CVI was k=3. These results were obtained from 13 CVI (40.63%) as illustrated in **Figure 3**.

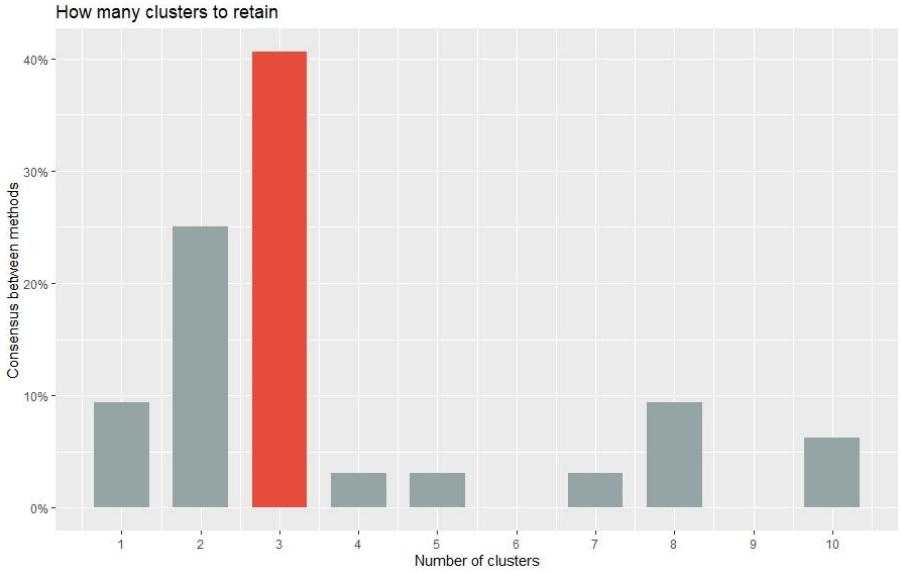

Figure 3. The Best Number of Clusters.

B. Clustering

The clustering results using the WM, SL, CL, AL, and DA methods are presented in **Table 2**. Most cluster members were in cluster 1 of the SL method, namely 32 provinces. However, some clusters had only one member, as in clusters 2 and 3 of the SL method (North Sumatra and Riau Islands, respectively), and cluster 3 of the AL method (Riau Islands). It can be an indication that objects in these clusters were outliers.

Table 2. Clustering for Each Method.

Cluster	WM	SL	CL	AL	DA
1	18	32	17	28	19
2	5	1	14	5	6
3	11	1	3	1	9

C. Cluster Validation

Table 3 shows that the *Sw*, AD, and FM values of WM outperformed other methods. However, SL was the best clustering approach based on *Di*, APN, and ADB values. Therefore, the cluster dendrograms of these two approaches were analyzed to determine the method that produced the most compact cluster.

Table 3. Cluster Validation for Five Approaches.

Method	Internal Validation		Stability Validation			
	Di	***Sw***	***APN***	***AD***	***ADB***	***FM***
WM	0.165353	0.316926	0.244756	2.031496	0.737580	0.945801
SL	0.349167	0.193483	0.056526	2.405375	0.281777	1.009646
CL	0.170509	0.258344	0.281801	2.171433	0.906324	0.979720
AL	0.134681	0.224161	0.121008	2.425087	0.745773	1.022827
DA	0.165353	0.288769	0.262728	2.100745	0.812899	0.992869

Based on the cluster dendrogram visualization, the WM dendrogram in **Figure 4** was clustered better (more compact) than the SL dendrogram in **Figure 5**. The WM dendrogram had a higher similarity value between objects in each cluster than SL. Therefore, WM was selected as the most appropriate method because clusters were more compact, the sensitivity level was low, and the dendrogram cluster was easier to interpret.

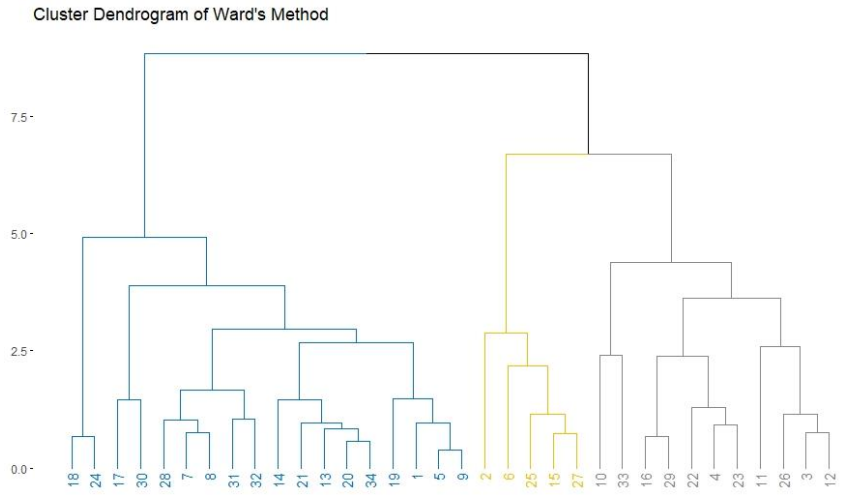

Figure 4. Cluster Dendrogram of Ward's Method.

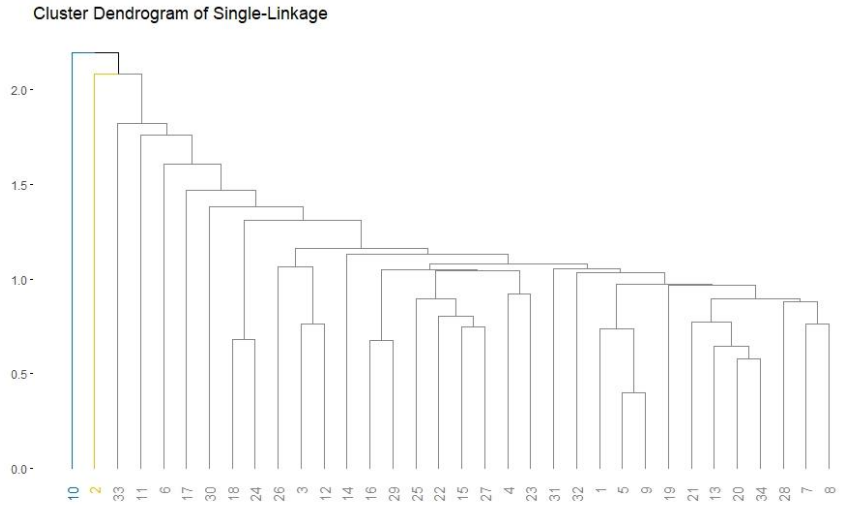

Figure 5. Cluster dendrogram of Single-Linkage.

Figure 6 presents SW visualization of the WM that shows that only one province in cluster 3 (South Kalimantan) has a negative value. It means that the province may be in the wrong cluster.

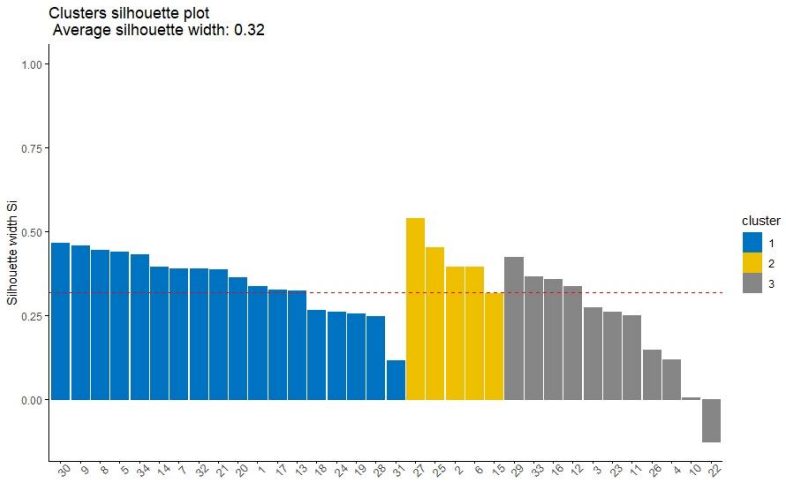

Figure 6. Average Silhouette Width of Ward's Method.

Based on **Table 1**, cluster 2 containing five provinces with the high 'murder' indicator, moderate 'victim' and 'safe' indicators, and the moderate-high 'report' indicator was a province cluster that was prone to crime. Provinces in this category should be prioritized for mitigation. Cluster 3 contained eleven provinces with the moderate 'murder', 'victim' and 'report' between medium-high and the low 'safe' was a province cluster that was less safe for people to walk alone. Finally, 18 provinces in cluster 1 with indicators of 'murder' and 'report' between low-medium, the 'victim' indicator between low-high, and the 'safe' indicator between medium-high was a province cluster that was relatively safe from crime. More clearly, **Table 4** presents the clustering results and is mapped in **Figure 7**.

Table 4. Province Clustering Using Ward's Method.

Province	Cluster
(1) Aceh, (5) Jambi, (7) Bengkulu, (8) Lampung, (9) Kep. Bangka Belitung, (13) Jawa Tengah, (14) DI Yogyakarta, (17) Bali, (18) Nusa Tenggara Barat, (19) Nusa Tenggara Timur, (20) Kalimantan Barat, (21) Kalimantan Tengah, (24) Kalimantan Utara, (28) Sulawesi Tenggara, (30) Sulawesi Barat, (31) Maluku, (32) Maluku Utara, (34) Papua	1
(2) Sumatera Utara, (6) Sumatera Selatan, (15) Jawa Timur, (25) Sulawesi Utara, (27) Sulawesi Selatan	2
(3) Sumatera Barat, (4) Riau, (10) Kep. Riau, (11) DKI Jakarta, (12) Jawa Barat, (16) Banten, (22) Kalimantan Selatan, (23) Kalimantan Timur, (26) Sulawesi Tengah, (29) Gorontalo, (33) Papua Barat	3

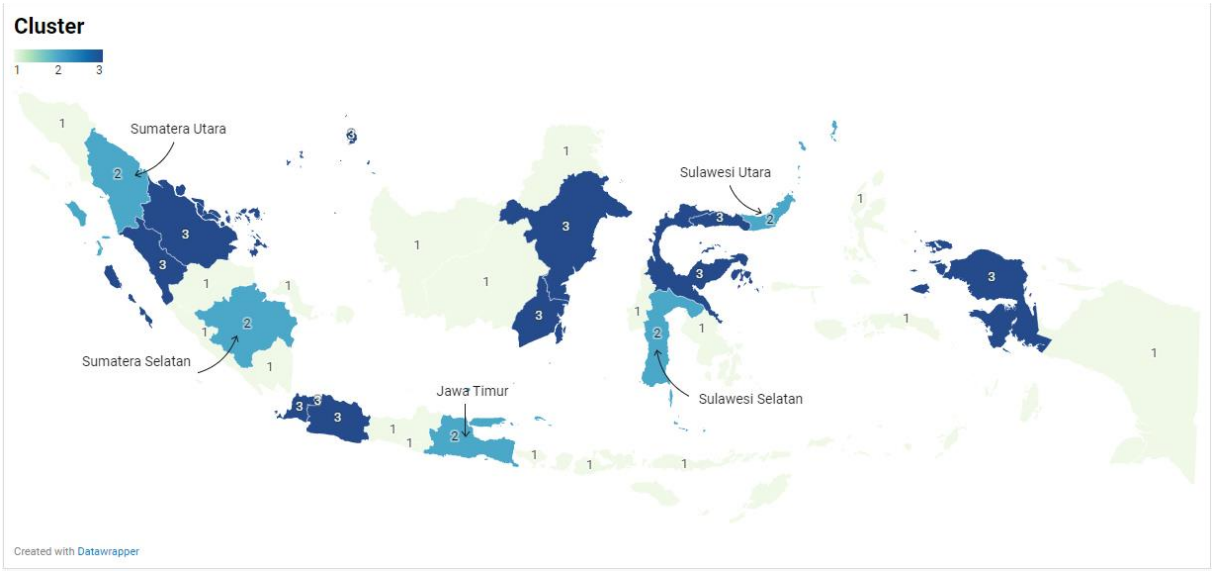

Figure 7. The Province Clustering Map.

Conclusion

This study introduces an approach to clustering provinces using five hierarchical clustering methods: Single-Linkage, Average-Linkage, Complete-Linkage, Ward, and Division Analysis. The clustering was based on four indicators from Goal 16 of the SDGs related to security and resulted in three clusters. Furthermore, the compactness of the clustering results was validated and compared to ß using six cluster validations.

Based on the validation results, it can be concluded that Ward's method outperforms other methods in compactness. Another finding is that around 15% of provinces in Indonesia (5 out of 34 provinces) are prone to crime, and more than 50% of provinces in Indonesia (18 out of 34) are relatively safe from crime.

The author plans to work using different clustering methods to obtain more optimal results. The analysis of the application of several other indicators that affect the clustering results will be a work in the future.

References

[1] r. n. fahmi, m. jajuli, and n. sulistiyowati, "Analisis pemetaan tingkat kriminalitas di kabupaten karawang menggunakan algoritma k-means," *intecoms j. inf. technol. comput. sci.*, vol. 4, no. 1, pp. 67–79, jun. 2021.

[2] a. joshi, a. s. sabitha, and t. choudhury, "Crime analysis using k-means clustering," in *2017 3rd international conference on computational intelligence and networks (cine)*, 2017, pp. 33–39.

[3] e. d. aritonang, h. s. tambunan, j. t. hardinata, e. irawan, and d. suhendro, "Penerapan data mining dalam

mengelompokkan provinsi rawan kejahatan menggunakan algoritma k-means," in *komik (konferensi nasional teknologi informasi dan komputer)*, 2020, pp. 35–42.

[4] a. a. alkhaibari and ping-tsai chung, "Cluster analysis for reducing city crime rates," in *2017 ieee long island systems, applications and technology conference (lisat)*, 2017, pp. 1–6.

[5] n. g. p. r. taram, i. k. g. sukarsa, and i. g. a. m. srinadi, "Pengelompokan tingkat kriminalitas dengan metode agglomerative dan k-means serta peubah pencirinya," *e-jurnal mat.*, vol. 8, no. 2, pp. 102–111, may 2019.

[6] e. h. s. atmaja, "Pengelompokan tingkat kriminalitas di kota yogyakarta dengan menggunakan metode k-means clustering," in *seminar nasional riset dan teknologi terapan*, 2018, pp. ik7–ik15.

[7] h. d. tampubolon, s. suhada, m. safii, s. solikhun, and d. suhendro, "Penerapan algoritma k-means dan k-medoids clustering untuk mengelompokkan tindak kriminalitas berdasarkan provinsi," *j. ilmu komput. dan teknol.*, vol. 2, no. 2, pp. 6–12, nov. 2021.

[8] b. sivanagaleela and s. rajesh, "Crime analysis and prediction using fuzzy c-means algorithm," in *2019 3rd international conference on trends in electronics and informatics (icoei)*, 2019, pp. 595–599.

[9] m. d. simatupang and a. w. wijayanto, "Analisis klaster berdasarkan tindakan kriminalitas di indonesia tahun 2019," *j. stat. ind. dan komputasi*, vol. 6, no. 1, pp. 10–19, 2021.

[10] s. k. appiah, k. wirekoh, e. n. aidoo, s. d. oduro, and y. d. arthur, "A model-based clustering of expectation–maximization and k -means algorithms in crime hotspot analysis," *res. math.*, vol. 9, no. 1, pp. 1–12, dec. 2022.

[11] n. azis *et al.*, "Mapping study using the unsupervised learning clustering approach," *iop conf. ser. mater. sci. eng.*, vol. 1088, no. 1, pp. 1–7, feb. 2021.

[12] m. a. nahdliyah, t. widiharih, and a. prahutama, "Metode k-medoids clustering dengan validasi silhouette index dan c-indeks (studi kasus jumlah kriminalitas kabupaten/kota di jawa tengah tahun 2018)," *j. gaussian*, vol. 8, no. 2, pp. 161–170, may 2019.

[13] a. baqir, s. ul rehman, s. malik, f. ul mustafa, and u. ahmad, "Evaluating the performance of hierarchical clustering algorithms to detect spatio-temporal crime hot-spots," in *2020 3rd international conference on computing, mathematics and engineering technologies (icomet)*, 2020, pp. 1–5.

[14] statistics-indonesia, "Social resilience," *social resilience*, 2022. [online]. available: https://bps.go.id/subject/34/politik-dan-keamanan.html.

[15] s. f. mu'afa and n. ulinnuha, "Perbandingan metode single linkage, complete linkage dan average linkage dalam pengelompokan kecamatan berdasarkan variabel jenis ternak kabupaten sidoarjo," *inf. j. ilm. bid. teknol. inf. dan komun.*, vol. 4, no. 2, pp. 1–5, jul. 2019.

[16] y. ogasawara and m. kon, "Two clustering methods based on the ward's method and dendrograms with interval-valued dissimilarities for interval-valued data," *int. j. approx. reason.*, vol. 129, pp. 103–121, feb. 2021.

[17] l. bellanger, a. coulon, and p. husi, "Determination of cultural areas based on medieval pottery using an original divisive hierarchical clustering method with geographical constraint (mapclust)," *j. archaeol. sci.*, vol. 132, pp. 1–18, aug. 2021.

[18] r. gustriansyah, n. suhandi, and f. antony, "Clustering optimization in rfm analysis based on k-means," *indones. j. electr. eng. comput. sci.*, vol. 18, no. 1, pp. 470–477, 2020.

[19] s. cahyani, r. wiryasaputra, and r. gustriansyah, "Identifikasi huruf kapital tulisan tangan menggunakan linear discriminant analysis dan euclidean distance," *j. sist. inf. bisnis*, vol. 8, no. 1, pp. 57–67, apr. 2018.

[20] f. batool and c. hennig, "clustering with the average silhouette width," *comput. stat. data anal.*, vol. 158, pp. 1–18, jun. 2021.

[21] s. liang, d. han, and y. yang, "Cluster validity index for irregular clustering results," *appl. soft comput.*, vol. 95, pp. 1–17, oct. 2020.

[22] x. zhang, m. zhao, j. appiah, and m. d. fontaine, "Improving interstate freeway travel time reliability analysis by clustering travel time distributions," *Transp. Res. Rec. J. Transp. Res. Board*, vol. 2675, no. 10, pp. 566–577, Oct. 2021.