khazanah informatika
Jurnal Ilmu Komputer dan Informatika

Silhouette Density Canopy K-Means for Mapping the Quality of Education Based on the Results of the 2019 National Exam in Banyumas Regency

Ridho Ananda
Faculty of Industrial and Informatics Engineering
Institut Teknologi Telkom Purwokerto
Indonesia
ridho@ittelkom-pwt.ac.id

Abstract-Mapping the quality of education units is needed by stakeholders in education. To do this, clustering is considered as one of the methods that can be applied. K-means is a popular algorithm in the clustering method. In its process, K-means requires initial centroids randomly. Some scientists have proposed algorithms to determine the number of initial centroids and their location, one of which is density canopy (DC) algorithm. In the process, DC forms centroids based on the number of neighbors. This study proposes additional Silhouette criteria for DC algorithm. The development of DC is called Silhouette Density Canopy (SDC). SDC K-means (SDCKM) is applied to map the quality of education units and is compared with DC K-means (DCKM) and K-means (KM). The data used in this study originated from the 2019 senior high school national examination dataset of natural science, social science, and language programs in the Banyumas Regency. The results of the study revealed that clustering through SDKCM was better than DCKM and KM, but it took more time in the process. Mapping the quality of education with SDKCM formed three clusters for social science and natural science datasets and two clusters for language program dataset. Schools included in cluster 2 had a better quality of education compared to other schools.

Keywords: Density canopy, K-means, Quality mapping, Silhouette**.**

1. Introduction

National Examination (UN) is a national-scale examination activity with the reference of graduate competence standard [1]. UN is held at the elementary level (SD), junior high school (SMP), and senior high school (SMA) or equivalent level in certain subjects. The government institution obliged to carry out the UN is Badan Standar Nasional Pendidikan (BNSP) aimed at measuring the fulfillment of graduate competence. The results of the UN then can be used as a mapping of the quality of education units [2].

The mapping of the quality of the education unit program is important to help education stakeholders in making education-related policies. BNSP has mapped the quality of the education unit program based on the UN results [1]. The quality is then classified into four criteria: (a) "excellent" if , (b) "good" if (c) "satisfactory" if , and (d) "poor" if . These criteria certainly have weaknesses when applied to the average UN results since the average calculation is sensitive to extreme values or outliers [3]. Therefore, we need a specific method that can provide the mapping of the quality of the education unit program concerning UN results where the calculation does not directly use the average score. Clustering is considered to solve this problem.

Clustering is a statistical classification technique to determine the classification of an individual of a population to be grouped into the same group or different group based on the quantitative comparison of the measured variables [4]. A clustering algorithm is needed to apply the clustering process. One of the classic and well-known algorithms is K-means, proposed by MacQueen[5]. K-means is included in the the unsupervised learning group category. The main principle of the K-means is classifying objects based on Euclidean distance measurement. The use of K-means is quite popular in the field of technology [6] [7], health [8], education [9][10][11] and other fields. In the process, K-means requires initial centroids in the first step. A centroid is the central data in certain cluster. This determination is often done randomly which affects the accuracy of the clustering results. To overcome this problem, some scientists have carried out some research and development of the K-means and proposed new algorithms such as canopy [12], K-means++ [13], K-means-u [14], and DCKM [15]. All these algorithms have been tested by [15] with the conclusion that DKCM is the most effective algorithm and can overcome the existence of extreme data or outliers. In DC, the first centroid is chosen from the object which has the most neighbors based on its Euclidean distance. This leads to the formation of less effective initial centroids due to the possibility that the first centroid is most likely not far from the data centroid and has the most members. Whereas the other centroids are obtained from the periphery of data that are not neighbors to the original centroid.

Based on the description above, research related to mapping the quality of the education unit program was conducted with the basis of the results of the UN. The data used in this study were data of 2019 senior high school UN results in Banyumas Regency. This current study used the DC algorithm as preprocessing of K-means for its ability to choose the initial centroid and its optimal location. On the other hand, the algorithm was appropriate since data were numerical. This study proposes a modification to the DC algorithm, which in the algorithm process the centroid determination should also consider Silhouette criteria. This study is expected to provide information for education stakeholders related to the mapping of the quality of the education unit program based on the results of the UN at the SMA level in Banyumas Regency in 2019. It is also hoped that the proposed modification can contribute to meaningful knowledge especially in clustering.

2. Research Method

a. Research Procedure

The research procedure for each dataset is shown in figure 1.

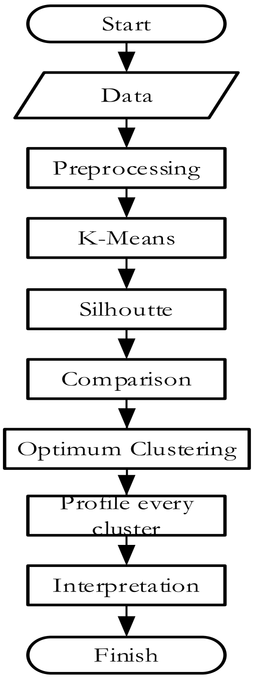


Figure 1. The Diagram of the Research Procedure

The computational process used Matlab 2014a software. In the preprocessing stage, this study used a modified density canopy and density canopy. Preprocessing is an initial step performed before the main step (clustering process). Then the clustering process was done with K-means and the Silhouette value was calculated. Furthermore, the clustering results of modified density canopy k-means were compared with the results of density canopy k-means and regular k-means. The optimum results will be used to map the quality of the education unit in Banyumas Regency.

b. Algoritme Density Canopy

Density canopy (DC) algorithm, proposed by [15], is the development of the Canopy algorithm [12]. The algorithm is a preprocessing of K-means to determine the initial centroids. DC selects the first centroid based on maximum density (number of neighbors) as shown in Figure 2. The figure provides an illustration of centroid selection on the DC algorithm. The steps of the DC algorithm are as follows:

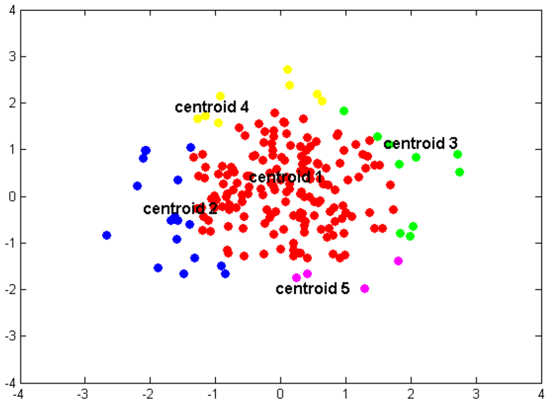


Figure 2. Distribution of initial centroids of the DC algorithm on random data with normal distribution

Step 1. Suppose matrix dataset sized The first step, calculate the average distance between objects with the formula (1) and the density of each object with the formula (2).

$$\bar{d}_E = \frac{2}{n(n-1)}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} d_E(\mathbf{x}_i,\mathbf{x}_j). \quad (1)$$

is the row vector of the object-i and is Euclidean distance of the object-i to object-j or vice versa using formula (9).

$$\rho(i) = \sum_{j=1}^{n} f\left(d_E(\mathbf{x}_i,\mathbf{x}_j) - \bar{d}_E\right) \quad (2)$$

where valued 1 for and for the rest. To each object-j that fulfills is considered as the closest neighbor of the object-i and the matrix of neighboring objects is formed as shown in equation (3).

$$\mathbf{N}_i = \left[\mathbf{x}_i, \mathbf{x}_{j_1}, \mathbf{x}_{j_2}, \cdots, \mathbf{x}_{j_{\rho(i)}}\right] \tag{3}$$

Step 2. Select object-i with maximum density (maximum) as the initial centroid, then objects which enter are deleted from the dataset.

Step 3. The rest of the objects on the dataset are calculated and the neighbors are determined or to each object-i. The next the calculation of the average distance between objects in the neighbor matrix of the object-i so that is obtained And then calculating local density with the formula (4).

$$q(i) = \begin{cases} \min d_E(\mathbf{x}_i, \mathbf{x}_j) & if\ \exists j \ni \rho(i) < \rho(j) \\ \square & \\ \text{maks}\, d_E(\mathbf{x}_i, \mathbf{x}_j) & if\ \forall j \ni \rho(i) > \rho(j) \end{cases} \tag{4}$$

Step 4. Calculate the weight of the object with the formula (5) for .

$$w(i) = \rho(i) \times \frac{1}{a(i)} \times q(i). \tag{5}$$

Step 5. Object with the highest score is chosen as the next centroid and objects in are deleted from the dataset.

Step 6. Repeat step 3 until no objects left in the dataset.

The obtained centroids are used as initial centroids in the grouping process with K-means.

From figure 2, it can be seen that the distribution of the centroids indicates unequal neighbor distribution. Centroids formed earlier have more neighbors than centroids come afterward. This is because DC is greedy in the process in which the centroid is chosen from the most neighbors consequently the next centroid is obtained from the rest of the dataset. There is a possibility that the last centroid has no neighbor because its neighbors have been claimed by previous centroids. Considering this condition, the current study proposes additional criteria to determine centroid with DC. The proposed additional criteria are inter-cluster distance and nearest cluster to the centroid candidate using the Silhouette formula.

c. Algoritme Silhouette Density Canopy

The underlying condition for the proposed additional criteria in the density canopy algorithm is the centroids selection based on the number of neighbors they have. The additional criteria proposed in this study are the Silhouette criteria. Figure 3 is the concept of the added criteria. The distance of a centroid to its neighbors is called intracluster distance, while the distance of centroid candidate to the objects that are not its neighbors is called the nearest cluster distance. The choice for Silhouette criteria is because it can determine many optimum clusters [16]. The best results of clustering are obtained when the Silhouette value is maximum. Based on this information, the multiplication operation will be used to combine the Silhouette value with maximum neighbor selection. Further, the proposed algorithm is named Silhouette density canopy (SDC). The step of the SDC are as follows:

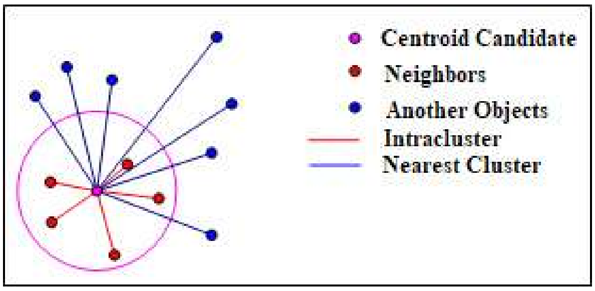


Figure 3. The Centroid cadidate based on Silhouette criteria on SDC

Step 1. Suppose matrix dataset sized The first step, calculate the distance between objects with the formula (1) and the density of each object with the formula (2). The next, create the neighbor matrix of the object-i with equation (3).

Step 2. Calculate the Silhouette criteria of the object-i with the formula (6).

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{6}$$

Where is the average distance of the object-i to its neighbors. is the average distance of from all other objects out of its neighbors.

Step 3. Select the object-i with maximum criteria as initial centroid then objects in are deleted from. Maximum criteria are obtained from formula (7).

$$\text{argmax}_i(p(i) \times s(i)) \tag{7}$$

Step 4. The ρ*(i)* of the rest of the objects on the dataset are calculated and their neighbors are determined or for each object-i. Then, calculate the average distance between objects on the neighbor matrix of the object-i and their Silhouette criteria . The next, calculate the local density with the formula (4).

Step 5. Calculate *w(i)* of each object with formula (8) for.

$$w(i) = \rho(i) \times s(i) \times \frac{1}{a(i)} \times q(i) \tag{8}$$

Step 6. The object with the highest score assigned as the next centroid and the objects in are deleted from the dataset.

Step 7. Repeat step 3 until no objects left in the dataset.

d. K-means

K-means (KM) algorithm is a grouping algorithm that minimizes the distance between the objects of the same group and maximizes dissimilarity between objects of different groups. The dissimilarity size used is the Eucladian distance [17] which is calculated with the formula (9).

The steps of the KM algorithm are as follows:

step 1. By using centroid obtained from Density canopy algorithm, for example, , every object is assigned to the group closest to its centroid.

step 2. Determine the new centroid of the average member. Every object is allocated to the group closest to the centroid formed. Each object can move to other groups if the new centroid makes it closer to the previous centroid.

step 3. Repeat step 2 iteratively until there are no changes in the grouping.

e. Model Validation

The validation is needed to measure the quality of clustering. There are two types of validation, those are external validation and internal validation [18]. In this study, the researchers only used internal validation. The external validation was not performed due to the absence of initial group information. It is said that the accuracy of the internal validation is higher than the external validation [18][19]. The internal validation used was the Silhouette index using formula (6). In the Silhouette validation, is the Silhouette size of the object-i. is the average distance between and other objects within a cluster (intracluster). is the average distance between and the objects in the nearest cluster. The higher the value the more appropriate the placement of the objects in the group. Silhouette index is obtained from the average size of the silhouette. Silhouette index ranges from -1 to 1, where the value closer to 1 indicates the object is well matched to its cluster [20]. The ability of the Silhouette index to validate the results of grouping is considered to be better than some validations in other fields [21]. The use of Silhouette validation has also been used, among others to validate clustering data of the automatic dependent surveillance-broadcast [22], clustering the province in Indonesia based on rice production [23], and the clustering of dengue-prone areas [24].

3. Results and Discussion

a. Data

The data used in this study are the report of UN results of the SMA level for natural science, social science, and language programs in Banyumas Regency in 2019. The data were obtained from the official report of the Ministry of Education and Culture. The data are in the form of matrix sized for natural science, for social science, and for a language program. Table 1 and table2 provide information on the list of high schools in the dataset.

Table 1. The list of high schools with a language program in dataset

Language Program	
Variable	Senior High School
B1	SMA N Ajibarang
B2	SMA N 1 Purwokerto
B3	SMA N 2 Purwokerto
B4	SMA N 5 Purwokerto

Table 2. The list of high schools with natural science and social science programs in dataset

Natural Science Program		Social Science Program	
Variable	SMA	Variable	SMA
A1	SMA N Ajibarang	S1	SMA N Ajibarang
A2	SMA N Banyumas	S2	SMA N Banyumas
A3	SMA N Baturraden	S3	SMA N Baturraden
A4	SMA N Jatilawang	S4	SMA N Jatilawang
A5	SMA N Patikraja	S5	SMA N Patikraja
A6	SMA N 1 Purwokerto	S6	SMA N 1 Purwokerto
A7	SMA N 2 Purwokerto	S7	SMA N 2 Purwokerto
A8	SMA N 3 Purwokerto	S8	SMA N 3 Purwokerto
A9	SMA N 4 Purwokerto	S9	SMA N 4 Purwokerto
A10	SMA N 5 Purwokerto	S10	SMA N 5 Purwokerto
A11	SMA N 1 Rawalo	S11	SMA N 1 Rawalo

Natural Science Program		Social Science Program	
Variable	SMA	Variable	SMA
A12	SMA N 1 Sokaraja	S12	SMA N 1 Sokaraja
A13	SMA N Sumpiuh	S13	SMA N Sumpiuh
A14	SMA N Wangon	S14	SMA N Wangon
A15	SMA Brunderan	S15	SMA Brunderan
A16	SMA Diponegoro Simpiuh	S16	SMA Budi Utomo Sokaraja
A17	SMA Ma'arif NU 1 Ajibarang	S17	SMA Diponegoro 1 Purwokerto
A18	SMA Ma'arif NU 1 Kemranjen	S18	SMA Jendral Sudirman
A19	SMA Ma'arif NU 1 Sokaraja	S19	SMA Karya Bakti Jatilawang
A20	SMA Muh. 1 Purwokerto	S20	SMA Ma'arif NU 1 Ajibarang
A21	SMA PGRI Gumelar	S21	SMA Ma'arif NU 1 Kemranjen
A22	SMA Yos Sudarso	S22	SMA Muh. 1 Purwokerto
A23	SMA Al Irsyad	S23	SMA Muh. Sokaraja
A24	SMA Muhammadiyah BSZ	S24	SMA Muh. Tambak
A25	SMA Islam Andalusia Kebasen	S25	SMA PGRI Gumelar
A26	SMA Nasional 3 BPH	S26	SMA Al Irsyad
A27	MAN 1 Banyumas	S27	SMA Muhammadiyah BSZ
A28	MAN 2 Banyumas	S28	SMA El-Madani Rawalo
A29	MAN 3 Banyumas	S29	SMA Islam Andalusia Kebasen
A30	MA Al-Ikhsan Beji	S30	SMA Nasional 3 BPH
A31	MA Ma'arif NU 1 Kemranjen	S31	MAN 1 Banyumas
A32	MA Miftahul Huda Rawalo	S32	MAN 2 Banyumas
A33	MA PPPI Miftahussalam	S33	MAN 3 Banyumas
A34	MA Wathoniyah Islamiyah	S34	MA Al-Ikhsan Beji
A35	MA Al-Falah Jatilawang	S35	MA Ma'arif NU 1 Kemranjen
A36	MA Ar-Ridlo Pekucen	S36	MA Muhammadiyah Purwokerto
A37	MA Ma'arif NU 1 Cilongok	S37	MA PPPI Miftahussalam
		S38	MA Wathoniyah Islamiyah
		S39	MA Ma'arif NU 1 Kebasen
		S40	MA Ar-Ridlo Pekucen
		S41	MA Al-Hidayah Purwojati

Table 3. The display of UN 2019 dataset

School	Indonesian	Englis	Math	Physics	Chemistry	Biology
SMA N Ajibarang	86.47	69.19	56.15	66.11	62.5	68.56
SMA N Banyumas	86.71	75.33	58.59	60.42	70.09	70.7
SMA N Baturraden	76.17	53.06	36.61	53.61	46.35	55.71
SMA N Jatilawang	85.46	66	54.01	58.3	68.75	66.67
…	…	…	…	…	…	…

The variables contained in the dataset of the results of the national examination of the natural science program are the results of examinations in Indonesian, English, Mathematics, Physics, Chemistry, and Biology. For the social science program, the variables are Indonesian, English, Mathematics, Economics, Sociology, and Geography. In the language program, those are Indonesian, English, Mathematics, Indonesian literature, Anthropology, and Indonesian language and literature. Table 3 illustrates the data used in the study

b. Preprocessing Results

In the preprocessing stage, the process is run from the SDC and DC algorithm. In the SDC algorithm, the Silhouette criterion value of each object will be observed based on the number of neighbors as shown in figure 4. The figure is the visualization of the Silhouette criterion value of ordered objects from the fewest to the most. Figure 4 provides information that the Silhouette criteria are not proportional to the number of neighbors they have. It can be seen that the object with the fewest neighbor does not necessarily has the lowest Silhouette criterion and vice versa. This indicates that Silhouette criteria will affect clustering results.

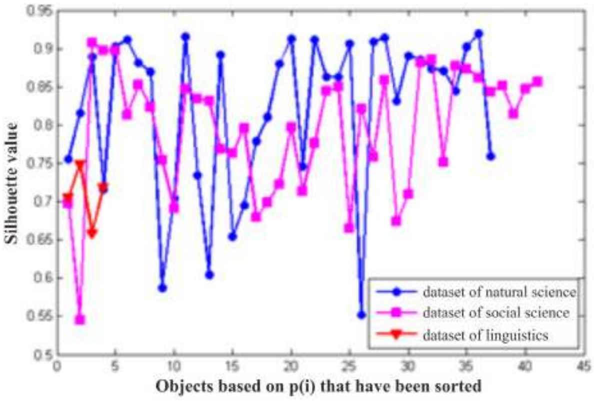


Figure 4. Silhouette criteria value based on ordered

Figure 5 is the visualization of the number of centroids and their location from the SDC and DC algorithm generated from principal component analysis(PCA, Principal Component Analysis). All visualizations have a component value of more than 88% based on the number of components 1 and 2. This indicates that visualization can maintain more than 88% of the information contained in the dataset so that it has relatively high quality. The location and the number of centroids in the natural science dataset obtained from the SDC algorithm (Figure 5a) and the DC algorithm (Figure 5d) have different values. It can be seen in the picture that SDC provides fewer initial centroids compared to DC. While in other datasets, the results of distribution and the number of initial centroids generated from SDC and DC are the same. The picture also provides information that SDC tends to choose most neighbors as initial centroids even though it has been offset by other criteria. Furthermore, the location and the number of initial centroids of the SDC and DC algorithms will be used as a prerequisite of the K-means algorithm.

c. Clustering Results

In the next stage, the clustering process is done by using K-means with SDC (SDCKM), K-means with DC (DCKM), and K-means without preprocessing algorithm (KM). Table 4 provides information on the number of clusters, Silhouette value, and the average time needed by each algorithm. The average time and Silhouette validation values are obtained from the iteration of each algorithm for 100 times in each dataset. The visualization of time and Silhouette validation of each iteration can be seen in figure 6.

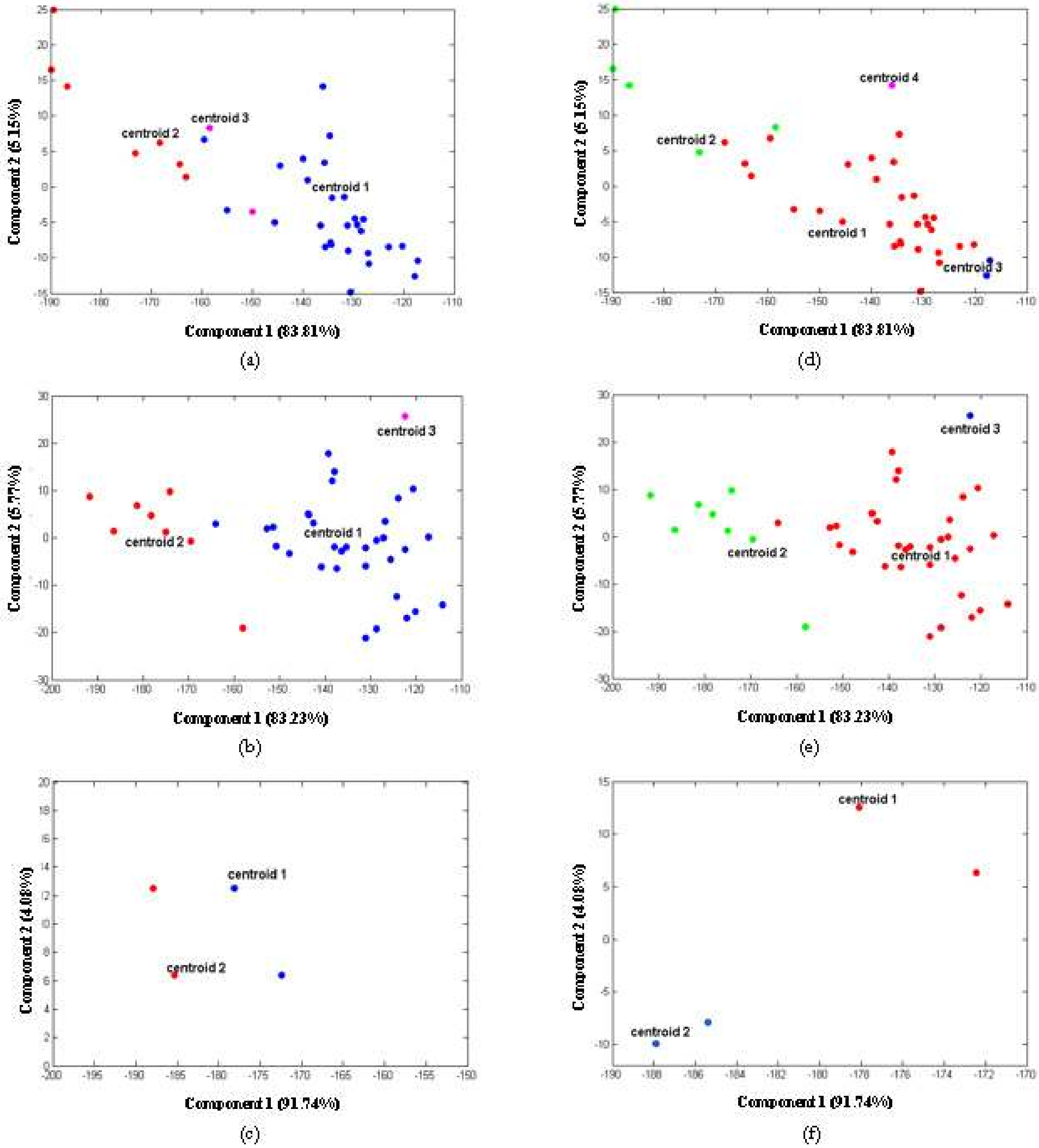
Table 4. The number of clusters, Silhouette value, and the average time of each algorithm

Dataset	Algorithm	Number of clusters	Silhouette	Time
Natural science	SDCKM	3	0.6895	6.938102
	DCKM	4	0.4475	1.189253
	KM	3	0.5013	0.485832
	KM	4	0.4148	0.622489
Social science	SDCKM	3	0.6340	8.629198
	DCKM	3	0.6340	0.962059
	KM	3	0.4472	0.600157
Language program	SDCKM	2	0.7079	0.083598
	DCKM	2	0.7079	0.075903
	KM	2	0.4303	0.048455

Table 4 shows that the SDCKM algorithm has a higher Silhouette value compared to other algorithms in the natural science dataset with three clusters formed.

Whereas in the social science and natural science dataset, SDCKM has the same value as the DCKM algorithm. Figures 6a, 6b, and 6c show that SDCKM and DCKM have a consistent Silhouette validation value for each iteration, it is known that this algorithm is deterministic. It is different when it comes to K-means where its validation values are not consistent since the determination of the initial centroids is done randomly. This creates a possibility that there is an opportunity element on the clustering results with K-means. The results of the Silhouette obtained by SDCKM in each dataset are optimum results with a Silhouette value higher than 0.5. Table 4 also shows the weakness of the SDCKM which is it needs the longest average time compared to other algorithms for each dataset, as shown in figure 6d, 6e, dan 6f.

d. Interpretation of Clustering Results



(a) (d)

(b) (e)

(c) (f)

Figure 5. Visualization of centroids distribution of the SDC algorithm in the dataset (a) natural science, (b) social science, and (c) language program and DC algorithm in dataset (d) natural science, (e) social science, and (f) language program.

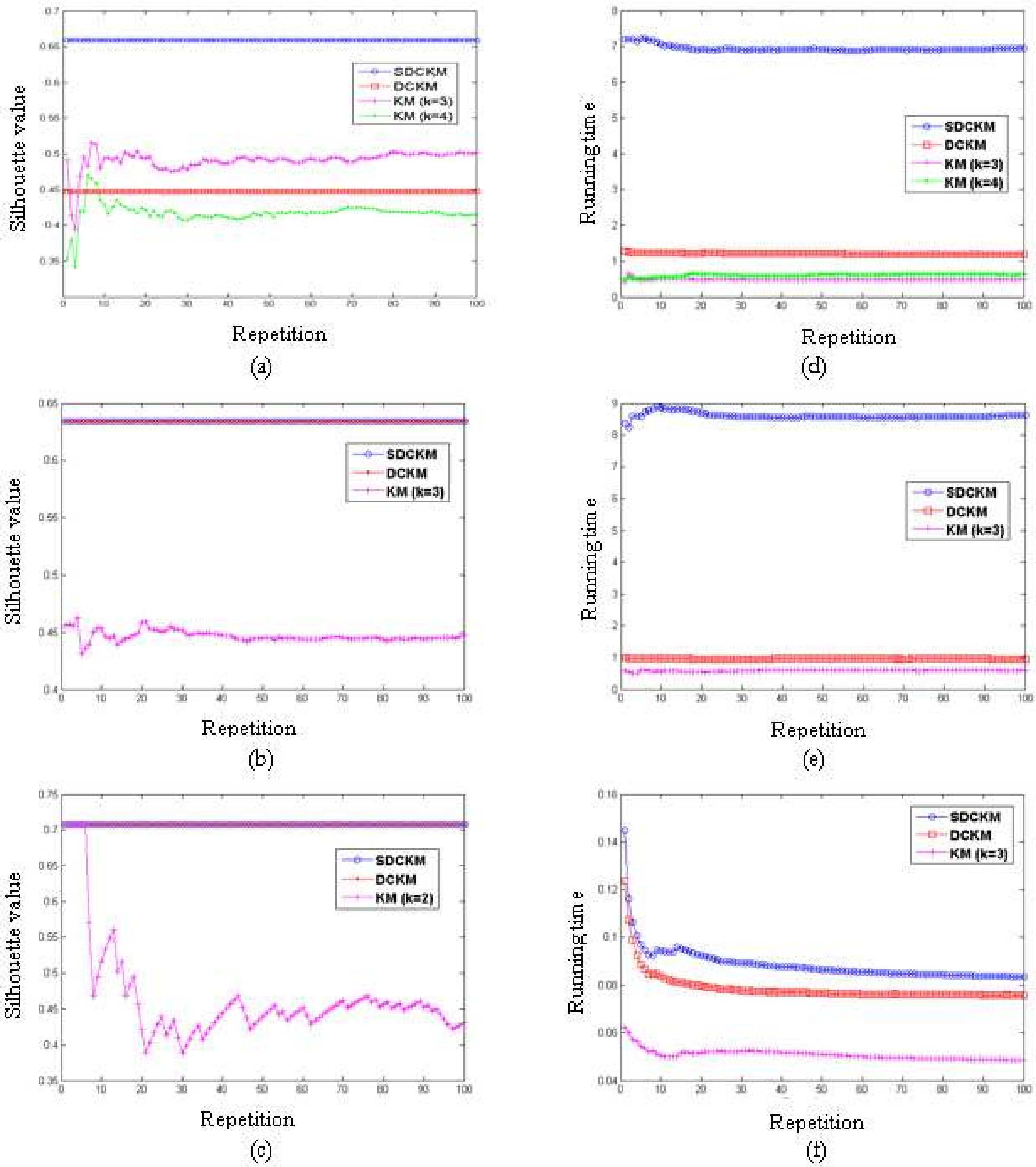


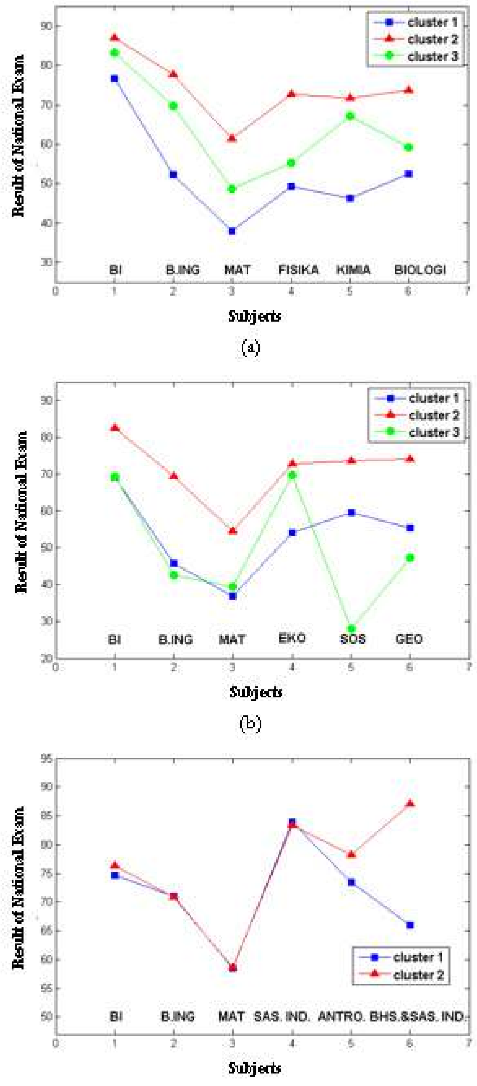
Figure 6. Visualization of the iterative Silhouette validation in the dataset (a) natural science, (b) social science, and (c) language program and running time in the dataset (d) natural science, (e) social science and (f) language program for each algorithm

Table 5. List of schools in certain cluster

Program	*Cluster* 1	*Cluster* 2	*Cluster* 3
Natural Science	A3, A5, A8, A11, A12, A14, A16, A17, A18, A19, A20, A21, A22, A24, A25, A27, A28, A29, A30, A31, A32, A33, A34, A35, A36, A37	A1, A2, AA6, A7, A23	A4, A9, A10, A13, A15, A26
Social Science	S3, S5, S8, S11, S12, S13, S14, S15, S16, S17, S18, S19, S20, S21, S22, S23, S24, S25, S27, S28, S29, S31, S32, S33, S34, S35, S36, S37, S38, S39, S41	S1, S2, S4, S6, S7, S9, S10, S26, S30	S40
Language	B2, B4	B1, B3	

Table 6. The comparison of the algorithms

Algorithm	Strength	Weakness
SDCKM	• It can determine the location and number of centroids optimally. • The determination of centroids considers 2 criteria which clustering results, those are the number of neighbors and Silhouette criteria.	• It requires a quite long time for the clustering process.
DCKM	• It can determine the location of centroids and the number of initial centroids • It requires a relatively short time.	• The determination of centroid is done based on the number of neighbors so that the formation of the initial centroid is not optimal.
KM	• It requires a relatively short time. • Simple algorithm	• The clustering results are not optimal since the determination of initial centroid and its number is done randomly.



(a)

(b)

(c)

Figure 7. The visualization of cluster profile for the dataset (a) natural science, (b) social science, and (c) language program

The interpretation of dataset mapping is done in the clustering process with the optimum result, which is the SDCKM result. The cluster's profile shown in figure 7. Figure 7a provides information on the cluster profile from the centroid of the clustering results on the natural science dataset. The results indicate that SMA in cluster 2 is better in quality compared to schools in cluster 1 and 3. This can be seen from the score obtained of the subjects tested, those are Indonesian (BI), English (B,ING), Mathematics (MAT), Physics, Chemistry, and Biology. The scores obtained by cluster 2 are higher than other clusters. while SMA in cluster 3 has better quality than schools in cluster 1. Figure 7b provides information that in social science dataset, the quality of SMA in cluster 2 is better than other schools for each subject such as Indonesian (BI), English (B. ING), Mathematics (MAT), Economics (EKO), Sociology (SOS), and Geography (GEO). Cluster 1 is considered to have relatively better quality than cluster 3since there are two subjects namely Sociology (SOS) and Geography (GEO) that show cluster 1 is better than cluster 3 with significant difference, on the contrary cluster 3 is only a way better than cluster 1 in one subject, that is Economics(EKO). For other subjects, cluster 1 and cluster 3 obtain relatively the same results. Figure 7c provides information that cluster 2 has a better quality of education unit than cluster 1 based on two subjects, namely Anthropology (ANTRO) and Indonesian language and literature (SAS. & BHS. IND.), these show that cluster 2 is better than cluster 1. Whereas in other subjects, such as Mathematics(MAT), Indonesian (BI), English (B.ING), and Indonesian literature (SAS. IND.), cluster 1 and cluster 2 relatively have the same results.

To find out the grouping of certain schools into a certain cluster, it can be seen in table 5 with information referring to table 1 and 2. From the table, it is obvious that for natural science schools included in cluster 2 are SMA N Ajibarang, SMA N Banyumas, SMA N 1 Purwokerto, SMA N 2 Purwokerto, and SMA Islam Teladan Al Irsyad Al Islamiyyah. Schools in cluster 3 are SMA N Jatilawang, SMA N 4 Purwokerto, SMA N 5 Purwokerto, SMA N Sumpiuh, SMA Bruderan Purwokerto, and SMA Nasional 3 Bahasa Putera Harapan. The rest of the schools which are not mentioned included in cluster 1. Meanwhile, in social science dataset, the schools included in in cluster 2 are SMA N Ajibarang, SMA N Banyumas, SMA N Jatilawang, SMA N 1 Purwokerto, SMA N 2 Purwokerto, SMA N 4 Purowkerto, and SMA N. The next, school included in cluster 3 is MA Ar- Ridlo Pekuncen only. Other schools that are not mentioned are in cluster 1. In the language program dataset, schools included in cluster 1 are SMA N 1 Purwokerto and SMA N 5 Purwokerto. While schools in cluster 2 are SMA Negeri Ajibarang and SMA N 2 Purwokerto. At the end of the discussion, table 6 displayed to provide information on the comparison of algorithms, such as SDCKM, DCKM, and KM, which are used in this study.

4. Conclusion

Based on the results and discussion presented in the previous parts, several conclusions are addressed. First, SDCKM has a better ability than CDKM in the clustering process of the 2019 UN dataset in Banyumas Regency. This can be seen from the Silhouette validation of the clustering results. Second, the time needed by the SDCKM algorithm is relatively long, so that it is not good enough to cluster a big dataset. Based on these conclusions, the proposed modification which includes Silhouette in the Density Canopy K-Means algorithm yield a better clustering result based on Silhouette validation. However, the addition of Silhouette criteria makes the clustering process to be longer than without the criteria. The results of the mapping of the quality of education concerning 2019 UN results in Banyumas Regency show that in the dataset of natural science, social science, and language program, the schools in cluster 2 have the best education quality compared to schools in cluster 1 and 3.

References

[1] Badan Standar Nasional Indonesia, “Prosedur operasional standar (POS) penyelenggaraan ujian nasional tahun pelajaran 2018/2019,” BNSP, indonesia, 2018.

[2] Badan Standar Nasional Indonesia, “Prosedur operasi standar ujian nasional sekolah menengah pertama, madrasah tsanawiyah, sekolah menengah pertama luar biasa, sekolah menengah atas, madrasah aliyah, sekolah menengah atas luar biasa, dan sekolah menengah kejuruan tahun pelajaran 2010/2011,” BNSP, Indonesia, 2011.

[3] A. Asra, Rudiansyah, *Statistika terapan untuk pembuat kebijakan dan pengambil keputusan*, 2nd Ed, In Media, 2014.

[4] A. J. Jain, “Data clustering: 50 years beyond k-means”, in *the 19th International Conference on Pattern Recognition, 1967.*

[5] J. B. MacQueen,” Some Methods for Classification of High Dimensional Data Sets with Application to Reference Matching”, in *Proceeding of 5th Berkeley Symposium on Mathematical Statistics and* Probability, 1967.

[6] D. S. Wardiani, N. Merlina, “Implementasi data mining untuk mengetahui manfaat RPTRA menggunakan metode k-means clustering,” *Jurnal PILAR Nusa Mandiri*, vol. 15, no. 1, pp. 125-132, 2019.

[7] A. A. Hassan, W. M. Shah, A. M. Husein, M. S. Talib, A. A. J. Mohammed, M. F. Iskandar, “Clustering approach in wireless sensor networks based on k-means: limitations and recommendations,” *IJRTE*,vol. 7, no. 6S5, pp. 119-126, 2019.

[8] U. Yelipe, S. Porika, M. Golla, “An Efficient Approach for Imputation and Classification of Medical Data Values Using Class-Based Clustering of Medical Records,” *Computers and Electrical Engineering*, vol. 66, pp. 487-504, 2018.

[9] Mardalius, “Implementasi algoritma k-means clustering untuk menentukan kelas kelompok bimbingan belajar tambahan (Studi kasus: siswa sma negeri 1 ranah pesisir)”, in *Proceding SEMILOKA ROYAL 2017 “Teknologi Mobile”*, 2017.

[10] H. Yuwafi,F. Marisa,I. D. Wijaya, “Implementasi Data Mining untuk Menentukan Santri Berprestasi di PP. Manarulhuda dengan Metode Clustering Algoritma K-means,” *Jurnal SPIRIT*,

vol. 11, no. 1, pp. 22-29, 2019.

[11] Y. S. Nugroho, S. N. Haryati, "Klasifikasi dan klastering penjurusan siswa sma negeri 3 boyolali," *Khazanah informatika*, vol. 1, no. 1, pp. 1-6, 2015.

[12] A. Kumar, Y. S. Ingle, A. Pande, P. Dhule, "Canopy clustering: a review on pre-clustering approach to k-means clustering," *IJIACS*, vol. 3, no. 5, pp. 22-29, 2014.

[13] J. Yoder, C. E. Priebe, "Semi-supervised K-means++," *The Journal of Statistical Computation and Simulation*, 2016.

[14] B. Fritzke, "The k-means-u* algorithm: non-local jumps and greedy retries improve k-means++ clustering[J]. 2017.

[15] G. Zhang, C. Zhang, H. Zhang, "Improved K-means Algorithm Based on Density Canopy," *Journal Knowledge-based Systems*, vol. 145, pp.289-297, 2018.

[16] A. R. Mamat, F. S. Mohamed, M. A. Mohamed, N. M. Rawi, M. I. Awang, "Silhouette index for determining optimal k-means clustering on images in different color models," *International Journal of Engineering & Technology*, vol. 7, no. 2, pp. 105-109, 2018.

[17] R. Johnson, D. W. Wichern, *Applied Multivariate Statistical Analysis*, 6th Ed, New York Pearson Education, 2007.

[18] E. Rendon, I. M. Abundez, C. Gutierrez, S. D. Zagal, A. Arizmendi, E. M. Quiroz, H. E. Arzate, "A comparison of internal and external cluster validation indexes," in *Proceeding of the 2011 American Conference on Applied Mathematics and the 5th WSEAS International Conference on Computer Engineering and Applications*, 2011.

[19] E. Rendon, I. Abudez, A. Arizmendi, E. M. Quiroz, "Internal Versus External Cluster Validation Indexes," *International Journal of Computers and Communications*, vol. 5, no. 1, pp. 27-34, 2011.

[20] L. Vendramin, R. J. G. B. Campello, E. R. Hruschka, "On the comparison of relative clustering validity criteria," in *Proceedings of the 2009 SIAM International Conference on Data Mining*, 2009.

[21] J. Baarsch, M. E. Celebi, "Investigation of Internal Validity Measures for K-means Clustering," in *Proceedings of the IMECS,* 2012.

[22] A. Saiful, J. L. Buliali,"Implementasi particle swarm optimization pada k-means untuk clustering data automatic dependent surveillance-broadcast," *Explora Informatika*, pp. 30-35, 2018.

[23] A. D. Munthe, "Penerapan clustering time series untuk menggerombolkan provinsi di Indonesia berdasarkan nilai produksi padi," *Jurnal Litbang Sukowati*, vol. 2, no. 2, pp. 1-11, 2019.

[24] Suprihatin, Y. R. W. Utami, D. Nugroho, "K-Means clustering untuk pemetaan daerah rawan demam berdarah," *Jurnal TIKomSIN*, vol. 7, no. 1, pp. 8-16, 20019.