Application of partitioning around medoids cluster for analysis of stunting in 100 priority regencies in Indonesia

M A Ramdani and S Abdullah*
Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Indonesia, Kampus Baru UI, Depok 16424, Indonesia

*sarini@sci.ui.ac.id

Abstract. In 2018, Indonesian government set 100 districts as the priority to reduce stunting. In this study, we hypothesize that the 100 determined districts should not be treated on an equal policy, due to some underlying factors that might affect the stunting in those districts. Thus, it is necessary to identify and analyze the grouping of 100 priority districts for stunting interventions in 2018 based on the National Team's indicators for the Acceleration of Poverty Reduction to see the severity of stunting. It is hoped that this clustering could be a reference for the government in determining priority regency groups to reduce stunting rates. Data on 100 districts, represented by eight numerical measurements and six categorical measurements were analyzed using Partitioning Around Medoids (PAM) method. Data similarity was measured using Gower distance, which can handle the clustering of mixed data types. We identified five priority district groups which provide meaningful insights. One of the groups has the worst stunting severity condition among other groups, for each indicator; implying the high priority to follow-up by the government. The majority of districts in Papua and East Nusa Tenggara Provinces are districts with poor stunting severity. We also found that poverty, proportion of the population with defecation in latrines, access to clean water and proper sanitation, number of integrated healthcare centers (posyandu) in a village, and number of doctors in each district are important factors that explain the stunting severity.

1. Introduction
The incidence of stunting for children under five in Indonesia is still relatively high compared to other countries. According to WHO in 2017, Indonesia is the third country with the highest stunting prevalence in the Southeast Asia region and is the third country with the highest average prevalence of children under five in 2005-2017 in Asia with an average of 36.4% [1]. The results of Basic Health Research (Riskesdas) in 2018 showed that the prevalence of stunting under five in Indonesia was 30.8% [1]. According to the Central Statistics Agency (BPS) in Indonesia, the prevalence of stunting in Indonesia is still high compared to other middle-income countries [2].

Based on these conditions, the National Team for the Acceleration of Poverty Reduction (TNP2K), which is directly under the auspices of the Vice President of the Republic of Indonesia, made a work program to maximize the reduction in stunting rates in Indonesia, namely determining priority areas in efforts to deal with stunting in Indonesia. In 2018, the government set 100 priority districts/cities in Indonesia to reduce stunting rates. The policy of determining the 100 priority areas for reducing stunting in 2018 is not necessarily a policy that only carried out for the 2018 fiscal year program but a policy that will continue to be carried out by the government at least 2021 [3].

TNP2K determines priority districts to reduce stunting rates using stunting prevalence data for children under five in 2013, so it does not show the condition of the severity of stunting in these areas
in the past few years. Thus, we assume that it will be ineffective if to reduce this stunting rate the government provides an evenly distributed allocation of the state budget, regional budget, and attention to the 100 priority districts without considering several other underlying factors affect the high or low stunting in these districts.

Therefore, based on existing data and the problems described, it is necessary to identify and analyze the grouping of 100 priority districts in Indonesia for stunting interventions in 2018 based on several indicators set by TNP2K to see the condition or situation of stunting in 100 districts priority by using stunting prevalence data in 2018 based on the results of Riskesdas 2018. This grouping expected to be a reference for the government in determining priority district groups to reduce stunting rates in Indonesia. Thus, the government can maximize the allocation of the state budget, regional budget, and can focus and synergize resources to overcome stunting in the priority groups obtained. Besides, from this grouping, it is hoped that local governments can make the right policies based on the severity of stunting for each district.

Several methods are available for grouping purposes, with the popular one is K-means clustering [4]. This method is based on Euclidean distance, which become a hindrance when dealing with categorical data. Another approach in clustering is that the similarity of groups is measured through the covariates’ distributions, known as Finite Mixture Model (FMM) [5]. This method was found to be satisfactory in identifying the underlying groups and providing insights on the identified groups [6-8]. FMM also allows for the dynamicity of the group assignment, in the sense that one object can move from one group to the other groups based on the posterior probability of group membership. However, this method requires assumption on distributions of each covariate, which not always can be easily determined in some real application.

Data on 100 priority districts for stunting interventions in 2018 represented by eight numerical measurements and six categorical measurements. Therefore, this research requires a unique clustering method that can handle grouping in mixed data. In this research, we propose Partitioning Around Medoids (PAM) method with Gower distance as a distance measure to handle mixed data types [9,10]. The PAM method has advantages over other methods; namely, it can produce a more robust cluster [11]. The PAM method is useful in handling relatively small datasets and high-dimensional data [9]. Maione et al. [12] conducted grouping on mixed data type social research data using the PAM method using Gower distance. Jung et al. [9] grouped data on mixed data type child murders in South Korea using the PAM method using Gower distance.

2. Methods

2.1. Gower distance

Gower distance designed for mixed data type processing, which calculated to measure the difference between two observations. A low difference (distance) value indicates that two observations are similar, and a significant difference (distance) value indicates that the two observations have very different values [9]. The concept of Gower distance is that for each data type, a suitable distance function used to measure the difference between the values in the variable and ensure that the resulting distance values are in the range 0 and 1. Then calculate the root of the average square of each value of this distance to get the distance between observations [13].

The following is the distance function used for each type of variable:

- Ordinal variables and numeric variables (interval and ratio): use range-normalized Manhattan distance, which is defined as follows:

  If the \( m \) th variable is an ordinal variable or a numeric variable, then the distance between two observations \( i, j \) on the \( m \) th variable is:

  \[
  d_m(i, j) = \frac{|x_{i,m} - x_{j,m}|}{R_m}
  \]

  where \( x_{i,m} \) is the value for the \( I \) th observation in the \( m \) th variable, and \( R_m \) is the range from the \( m \) th variable.
• Nominal variables: use dice coefficient, which is defined as follows:

\[ d_m(i, j) = \begin{cases} 
0, & x_{i,m} = x_{j,m} \\
1, & x_{i,m} \neq x_{j,m} 
\end{cases} \tag{2} \]

where \( x_{i,m} \) is the value for the \( i \)th observation in the \( m \)th variable.

After calculating the distance for each variable, the overall distance between the two observations \( i, j \) and \( p \) of the variable can be calculated based on the distance value for each variable, which is defined as follows:

\[ d(i, j) = \sqrt{\frac{\sum_{m=1}^{p} w_m (d_m(i, j))^2}{\sum_{m=1}^{p} w_m}} \tag{3} \]

where \( w_m \) is the weight for the \( m \)th variable, which has a value of 1 if the two observations in the \( m \)th variable are known (not missing value), and 0 otherwise [10].

2.2. Partitioning Around Medoids

The Partitioning Around Medoids (PAM) methods is a part of partitioning clustering methods that groups a set of \( n \) observations into a set of \( k \) clusters. The PAM method based on finding a representative observation from a set of observations to be made cluster representative, these cluster representative are called medoids. After finding \( k \) representative observations from a set of observations, \( k \) clusters will be built by assigning each existing observation in the dataset to its closest representative observation [11]. The PAM cluster method is an iterative clustering procedure which has the following steps:

1. Select \( k \) observations randomly to be the medoids of the \( k \) clusters.
2. For each observation from the remaining \( n-k \) observations (other than the medoids selected in step one), calculate the distance from the observation to each medoid (which in this study calculates the distance using the Gower distance). Medoids, who has the lowest distance from these observations will take these observations as members of their clusters.
3. For each cluster, identify new candidate medoids, namely observations in a cluster that produce the lowest average distance between each member of a cluster and the new medoids. Medoid change will occur if there are observations other than the current medoids that produce the lowest average distance, which means that the average distance of each member of a cluster with these observations is lower than the average distance of each member of a cluster to the current medoids
4. If there is at least one cluster that changes medoids, as mentioned in step three, repeat step two. If there is no change in medoids in all clusters, then the PAM method is complete [13].

In Figure 1 below, an illustration of the stages of the PAM method is presented.
The selection of data points as medoids as cluster representatives causes the PAM method to be more robust compared to clustering methods that use cluster averages as cluster representative, such as $K$-Means clustering [11]. The use of the cluster average as a representative cluster will be very sensitive to outliers because the nature of the mean is quite sensitive to outliers.

2.3. Silhouette coefficient
In this study, we used the Silhouette coefficient method to estimate the optimal number of clusters in the dataset grouping. The Silhouette coefficient method used to ensure that the dataset’s observations are in the right cluster or the appropriate cluster [14]. The best number of clusters formed could be estimated by calculating the average of the Silhouette coefficient designed for a variety of a different number of clusters. All Silhouette coefficient values from various number of clusters visualized in one plot, which can show the quality of the number of clusters [12].

In calculating the Silhouette coefficient value for each observation ($S_i$), there are two important components, namely $d_{avg}^{in}$ and $d_{min}^{out}$ [15]. The value of $d_{avg}^{in}$ is the average distance of the $i$ th observation to all other observations in the same cluster. The value of $d_{avg}^{in}$ shows how much the $i$ th observation is not similar to all other observations in the same cluster. The smaller value of $d_{avg}^{in}$ indicates the more precise the observation is in a cluster [15].

Suppose $x_i$ is the $i$ th observation in the cluster $C_r$. So, the value of $d_{avg}^{in}$ can be calculated by Equation (4), where $|C_r|$ is the number of members of the cluster $C_r$

$$d_{avg}^{in} = \frac{1}{|C_r| - 1} \sum_{(j|x_j \in C_r,j \neq i)} d(i,j). \quad (4)$$

Meanwhile, $d_{min}^{out}$ is the average minimum distance between the $i$ th observation and all other observations that are not in the same cluster with the $i$ th observation. In other words, $d_{min}^{out}$ is the average distance between the $i$ th observation and the nearest cluster. The higher value of $d_{min}^{out}$ shows, the higher the difference between the $i$ th observation and other observations that are not in the same cluster with the $i$ th observation [15].

The value of $d_{min}^{out}$ can be calculated in two steps. The first stage is to calculate the average distance between the $i$ th observation and other observations in a different cluster from the cluster where
the \( i \) th observation located, which is denoted by \( d_{i}^{\text{out,s}} \). This calculation is done for each existing cluster. The next step is to calculate \( d_{\text{min}}^{\text{out}} \), which is the minimum of \( d_{i}^{\text{out,s}} \), previously calculated.

Suppose there is cluster \( C_s \), where \( s \neq r, s \in \{1,2,3,\ldots,k\} \). So, the value of \( d_{i}^{\text{out,s}} \) can be calculated by Equation (5) and the value of \( d_{\text{min}}^{\text{out}} \) can be calculated by Equation (6), where \( |C_s| \) is the number of members of the cluster \( C_s \).

\[
d_{i}^{\text{out,s}} = \frac{1}{|C_s|} \sum_{j \in C_s} d(i,j)
\]

(5)

\[
d_{\text{min}}^{\text{out}} = \min(d_{i}^{\text{out,s}} | s = 1,2,3,\ldots,k, s \neq r).
\]

(6)

After calculating the values of \( d_{\text{avg}}^{\text{in}} \) and \( d_{\text{min}}^{\text{out}} \), the Silhouette coefficient value in the \( i \) th observation in cluster \( C_r \) denoted by \( S_i \) can be calculated using Equation (7), namely:

\[
S_i = \frac{d_{\text{min}}^{\text{out}} - d_{\text{avg}}^{\text{in}}}{\text{Maximum}(d_{\text{avg}}^{\text{in}}, d_{\text{min}}^{\text{out}})}.
\]

(7)

The Silhouette coefficient value is in the range -1 to 1. A positive Silhouette coefficient value \( (d_{\text{min}}^{\text{out}} > d_{\text{avg}}^{\text{in}}) \) means that observation tends to be placed in the right cluster, while a negative Silhouette coefficient value \( (d_{\text{min}}^{\text{out}} < d_{\text{avg}}^{\text{in}}) \) interpret that observation tends to be placed in the wrong cluster [14].

To estimate the optimal number of clusters, we used the Global silhouette coefficient value \( (SC_g) \). The Global silhouette coefficient value can be used to measure how good the overall cluster results are, which can be calculated by Equation (8) below, where \( n \) is the total number of observations [15].

\[
SC_g = \frac{1}{n} \sum_{i=1}^{n} S_i.
\]

(8)

The most optimal number of clusters chosen is the number of clusters with a high Global silhouette coefficient value and produce clusters that have entirely different profiles or characteristics between the clusters formed [14].

2.4. Study Design

In this study, the steps to identify and analyze the grouping of 100 priority districts in Indonesia for stunting interventions in 2018 summarized in Figure 2 below:
3. Result and Discussion

3.1. Data
The data we use in this study is secondary data from 100 priority districts for stunting intervention in 2018. We use 14 variables in this study, which consists of eight numerical variables and six categorical variables. The variables of numerical type are the prevalence of stunting, the population, the poverty level, the proportion of the population with defecating behavior in the latrine, the proportion of households that have access to clean water, the proportion of households that have access to proper sanitation, the proportion of APBD spending on health functions, and the proportion of APBD expenditure on housing and public facilities for each district/city. Meanwhile, the variables of a categorical type are the availability of health centers, the availability of hospitals, the proportion of villages with a sufficient number of posyandu per village, the availability of doctors, the availability of midwives, and the availability of nurses for each district/city.

Data from 100 priority districts/cities for stunting interventions in 2018 were obtained from the TNP2K website (www.tnp2k.go.id), the 2018 Public Health Development Index (IPKM) book from Ministry of Health of the Republic of Indonesia, the Central Statistics Agency (BPS) in the 2019 figures for each province in Indonesia and the BPS website for the 2018 poverty level. In addition, there are also data on the longitude and latitude of 100 priority districts for stunting intervention in 2018, gathered from tanahair.indonesia.go.id.

3.2. The most optimal number of clusters
Based on section Silhouette coefficient above, in this section we will determine the most optimal number of clusters using RStudio software with a daisy package from the website https://cran.r-project.org/web/packages/cluster/cluster.pdf.

Check on Figure 3 below, it can be seen that the number of clusters (k) that has the highest Global silhouette coefficient value is $k = 2$, which is 0.4305536. Also, several clusters have high Global silhouette coefficient values, namely $k = 3$, $k = 5$, and $k = 4$, with the global silhouette coefficient values being 0.393817, 0.3907505, and 0.3781573, respectively. Thus, a cluster will be constructed using the number of clusters with a high Global silhouette coefficient value and seen whether to produce clusters that provide meaningful insights. It found that constructing clusters using the number of clusters of $k = 5$ resulted in clusters that provide meaningful insights so that the most optimal number of clusters chosen is $k = 5$. Thus, in this study the number of clusters to be constructed using the PAM method is $k = 5$. 
3.3. Distribution of districts for each cluster

In this section, we will describe a map of Indonesia and the 100 priority districts for stunting intervention in 2018 to see the distribution of districts for each cluster. Maps created using RStudio software with a leaflet packages (https://cran.r-project.org/web/packages/leaflet/leaflet.pdf).

Figure 4. Distribution of districts for each cluster

In Figure 4, the areas in purple indicate districts that are members of Cluster 1, areas in yellow indicate districts that are members of Cluster 2, areas in green indicate districts that are members of Cluster 3, areas that are in red (red salmon) indicates districts that are members of Cluster 4, and areas in black indicate districts that are members of Cluster 5. Based on the distribution of colours, it can be seen that the districts that are members of clusters 1, 2, and 4 come from regions various. It can also be seen that the districts that are members of cluster 3 are dominated by districts originating from Java Island. Meanwhile, the majority of districts originating from NTT and Papua Provinces are members of Cluster 2, Cluster 4, and Cluster 5.

3.4. Characteristics of each cluster

In this study, to facilitate the interpretation or determination of each cluster’s characteristics, the variables with the categorical type used the mode threshold value [16]. A category is said to be mode if it has a percentage that is greater than or equal to the mode threshold. The percentage of each category is the percentage of the number of districts in each category of a variable to the total number of districts. Category classified as the mode is a profile or characteristic of a variable [16]. For example, if the
number of categories for a variable is five, then the mode threshold is 20%. Thus, the cluster profile on these variables is explained by the category, which has a percentage of the number of members greater than or equal to 20%. Meanwhile, to facilitate the interpretation or determination of each cluster's characteristics for numeric variables, the average value of each cluster is used for each numeric variable.

Based on the above description, we obtain the characteristics of each cluster as the following. Cluster 1 is a cluster with the best average stunting prevalence and average poverty rates compared to other clusters. Cluster 1 has an average proportion of the population with defecating behavior in the latrine, the proportion of households with access to clean water, and a relatively high proportion of households with access to proper sanitation. Cluster 1 has the smallest average population compared to other clusters. Cluster 1 has a relatively more than a sufficient number of hospitals, health centers, doctors, midwives, and nurses. Cluster 1 has a relatively good number of posyandu per village. In general, cluster 1 is the best cluster when compared to other clusters.

Cluster 2 is a cluster that has a relatively high average prevalence rate of stunting. Cluster 2 has a relatively low average proportion of households that have access to proper sanitation, has a relatively good average proportion of households that have access to clean water, and has an average proportion of the population with defecating behavior in a latrine relatively high enough. Cluster 2 has the availability of health service facilities such as hospitals and health centers (puskesmas), similar to cluster 1. It has a relatively low number of doctors, has a relatively sufficient number of midwives, and has a relatively more than an adequate number of nurses. Cluster 2 has a relatively sufficient number of posyandu per village. In general, cluster 2 is in relatively better condition than cluster 4 and cluster 5.

Cluster 3 is a cluster that has an average stunting prevalence rate of 33.45%. Cluster 3 has a relatively low poverty rate. It has an average proportion of the population with defecating behavior in the latrine, the proportion of households that have access to clean water, and the proportion of households that have access to proper sanitation, which is relatively similar to cluster 1. Cluster 3 has the highest average population size compared to the cluster other. Cluster 3 has the highest average proportion of health function APBD spending compared to other clusters. Cluster 3 has hospital availability that is relatively similar to clusters 1 and 2. It has relatively less availability of health centers (puskesmas). It has relatively low availability of health personnel such as doctors, midwives, and nurses. Cluster 3 has a relatively good number of posyandu per village. In general, cluster 3 is in relatively better condition compared to clusters 2, 4 and 5.

Cluster 4 is a cluster with a relatively high average prevalence rate of stunting, such as cluster 2. Cluster 4 has an average proportion of the population with defecating behavior in the latrine and a relatively high proportion of households with access to clean water. Cluster 4 has an average proportion of households that have access to proper sanitation is similar to Cluster 2. Cluster 4 has the availability of hospitals and health centers that are relatively similar to cluster 1. Cluster 4 has the availability of health personnel such as doctors and nurses, which is similar to Cluster 3. relatively mixed, that is, less and sufficient. Cluster 4 has a relatively low number of posyandu per village. In general, Cluster 4 is in relatively better condition than cluster 5.

Finally, Cluster 5 is a cluster that has the highest average stunting prevalence rate and average poverty rate compared to other clusters. Cluster 5 has the average proportion of the population with defecating behavior in the latrine, the proportion of households that have access to clean water, and the proportion of households that have access to proper sanitation, which is relatively low compared to other clusters. It has the lowest average proportion of health function APBD spending compared to other clusters. Cluster 5 has relatively sufficient hospital availability and has a relatively similar availability of puskesmas to clusters 1, 2, and 4. It has the availability of health personnel such as doctors, midwives, and nurses that are relatively similar to Cluster 3. This cluster is also lacking the availability of the number of posyandu. In general, Cluster 5 has the worst condition compared to other clusters.

4. Conclusion
Based on the analysis results, several recommendations to the government can be made regarding the order of priority district clusters that require more attention. It can be concluded that the cluster that is the government's priority in the efforts to deal with stunting is Cluster 5. The next priority is Cluster 4, Cluster 2, Cluster 3, and the least priority is Cluster 1. It can also be concluded that of the 14 indicators
used in this study, several indicators appear to be associated with high or low stunting prevalence rates in a district. These indicators are the level of poverty, the proportion of the population with defecating behavior in the latrine, the proportion of households that have access to clean water, the proportion of households that have access to proper sanitation, the proportion of villages with a sufficient number of posyandu per village and the availability of doctors. The clusters whose districts have a high poverty rate also have a high prevalence rate of stunting. Thus, to further increase efforts to reduce stunting rates in the 100 priority districts, the government needs to pay more attention to these indicators. Based on the distribution of districts in each cluster, it can be seen that there is a gap in the condition of the severity of stunting in several districts in Indonesia. It can be seen that districts originating from the western part of Indonesia tend to do better compared to districts originating from the central and eastern parts of Indonesia. So, the government needs to pay more attention to districts originating from the central and eastern parts of Indonesia in the efforts to deal with stunting.

Acknowledgments

This research was funded by Directorate of Research and Development of Universitas Indonesia (DRPM UI) as a grant of Publikasi Terindeks Internasional (PUTI) Prosiding 2020 No. NKB-977/UN2.RST/HKP.05.00/2020. Authors wishing to acknowledge assistance or encouragement from colleagues, special work by technical staff, and financial support from Department of Mathematics, Faculty of Mathematics and Natural Sciences, University of Indonesia.

References

[1] Kementerian Kesehatan Republik Indonesia 2018 Situasi Balita Pendek (Stunting) di Indonesia Kementerian Kesehatan Republik Indonesia (Jakarta: Kementerian Kesehatan RI)

[2] Badan Pusat Statistik Provinsi Papua 2019 Workshop Enumerator/Training Center Studi Status Gizi Balita Indonesia (Provinsi Papua: BPS Provinsi Papua)

[3] Tim Nasional Percepatan Penanggulangan Kemiskinan (TNP2K) 2018 100 Kabupaten/Kota Prioritas untuk Intervensi Anak Kerdil (Stunting) Tim Nasional Percepatan Penanggulangan Kemiskinan (Jakarta: Tim Nasional Percepatan Penanggulangan Kemiskinan)

[4] MacQueen J 1967 methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability Jun 21 1 14 p 281-297

[5] McLachlan G J, Lee S X, Rathnayake S I 2019 Finite mixture models Annual review of statistics and its application 7 6 p 355-78

[6] Dwimantara N, Abdullah S, Bustamam A, Rachman A 2020 Implementation of Bayesian Mixture Models in identifying subpopulation of breast cancer patients based on blood test measurements In Journal of Physics: Conference Series 1494 p 012012

[7] Abdullah S, McGree J, White N, Mengersen K, Kerr G 2019 Profile regression for subgrouping patients with early stage Parkinson's disease arXiv preprint arXiv:1910.01864

[8] Sarini S Statistical methods for modelling falls and symptoms progression in patients with early stages of Parkinsons (Doctoral dissertation, Queensland University of Technology).

[9] Jung K H, Kim H, Lee E, Choi I, Lim H, Lee B, Choi B, Kim J, Kim H, and Hong H G 2020 Cluster analysis of child homicide in South Korea Child Abuse and Neglect 101 104322

[10] Myatt G J and Johnson W P 2008 Making Sense of Data II: A Practical Guide to Data Visualization, Advanced Data Mining Methods, and Applications (New Jersey: John Wiley & Son, Inc)

[11] Kaufman, L and Rousseuw, P J 2005 Finding Group in Data: An Introduction to Cluster Analysis (New Jersey: John Wiley & Son, Inc)

[12] Maione C, Nelson D R, and Barbosa R M 2019 Research on social data by means of cluster analysis Applied Computing and Informatics 15 153-62

[13] Martin, D 2016 Clustering Mixed Data Types in R (Wicked Good Data)

[14] Dos Santos T R L and Zárate L E 2015 Categorical data clustering: What similarity measure to recommend? Expert Systems with Applications 42 1247-60
[15] Pradana, C C B 2019 Pengelompokan Data Evaluasi Pembelajaran Menggunakan Algoritma K-Means++ Clustering Skripsi (Yogyakarta: Universitas Sanata Dharma Yogyakarta)

[16] Putri, M A 2020 Analisis Pengelompokan Sekolah Menengah Atas (SMA) di Indonesia Berdasarkan Fasilitas Pendidikan Skripsi (Depok: Universitas Indonesia)