Xplore (2021), Vol. 10, No. 1, 12–26
DOI: https://doi.org/10.29244/xplore.v10i1.272

Xplore: Journal of Statistics
Available at https://journal.stats.id/index.php/xplore
e-ISSN: 2655-2744

Two Step Cluster Application to Classify Villages in Kabupaten Madiun Based on Village Potential Data*

Alif Supandi[1], Asep Saefuddin[2‡], and Itasia Dina Sulvianti[3]

[1,2,3]Department of Statistics, IPB University, Indonesia
[‡]corresponding author: asaefuddin@apps.ipb.ac.id

Abstract

Village development is a fundamental part of national development. Developing villages requires information on society necessities. This research aims at clustering villages in Kabupaten Madiun which has similar characteristics among each other and identify characteristics of the built clusters. Therefore, specific problems in the clusters of villages may become the foundation to implement development. The method that used for grouping objects with combined variables is two-step cluster. This analysis was used 14 variables consist of six categorical variables and eight numerical variables. The clustering analysis produces four clusters. The clusters that need more attention to be developed was Cluster 2 which had minimum facilities and resources. The average Silhouette Coefficient for the clusters built was 0.3 which can be considered as fair value.

Keywords: cluster; silhouette coefficients; two step cluster.

*Received: Apr 2019; Reviewed: Des 2020; Published: Jan 2021

1 Introduction

1.1 Background

Village development is determined as an essential part of National Development. The core of village development is to develop and process the potential of the village that is implemented by the village community with the assistance of the relevant government so the results of the development can improve the welfare of the village community. As stated in the Minister of Home Affairs Regulation No. 114 of 2014 concerning Guidelines for Village Development Article 1 point 10, Village development planning is a process of stages of activities organized by the village government by involving Village Consultation Agency and community element's participation to utilize and allocate village resources in order to achieve village development goals.

Central Bureau of Statistics (BPS) of Kabupaten Madiun has recorded village potential for every village in Kabupaten Madiun in 2017. Village potential data consists of facilities and resources from each village, and it can be used as a measurement of the village development. In the beginning process of village development, this research aimed to provide information about the villages that has similar characteristics (based on village potential data) of each other. Therefore, the village community of Kabupaten Madiun can be focused to which potential characteristic needs more attention to be developed and what kind of development they most needed.

Clustering is a process for grouping objects by its similar characteristics among each other. Nowadays, the most common clustering method that is used is the hierarchical and non-hierarchical method. Those methods focus on data which has numerical data type (such as village population, number of worshipping facility, farming area, etc.), but BPS's village potential data is not only containing numerical data type but also mixed with categorical data type (such as village status, existence of art market, type of longest road, etc.). This kind of data needs special handling to ensemble between those categorical and numerical variables. Chiu *et al.* (2001) proposed and stated that Two Step Clustering can be used to cluster big objects with mixed both categorical and numerical variables. In the previous research, Putri (2005) and Alam & Ambarwati (2017) have conducted similar research that concerns on clustering villages based on village potential data.

1.2 Objectives

This research goals are to:

1. Determine the characteristics of *desa* (villages) and *kelurahan* (formal villages) in Madiun districts.
2. Clusterize the villages and formal villages in Madiun district based on their relatively similar potential characteristics.
3. Determine the general characteristics of the village clusters that has been built.

2 Methodology

2.1 Two Step Cluster Analysis

Cluster analysis is a multivariate analysis that can be used to cluster objects based on its similarity, so that those objects within the cluster have high similarity level, while those

objects in a different clusters have low similarity level Mattjik & Sumertajaya (2011).

There are two steps needed to perform Two Step Clustering (Chiu *et al.* (2001)), those are:

1. Pre-Clustering Stage
   This stage has been done with the constructing of Cluster Feature (CF) Tree that consists of some nodes and branch that can have leaf entries in each branches. The leaf entries represents the built sub-clusters to evaluate the new entry by examining their distance with it. The input object observed one by one rapidly and classified into the existed leaf entry or formed a new leaf entry based on the distance measured. If the distance is near, then the object gathered in the first-built leaf entry, if the distance is far, then the object formed a new leaf entry. This approximation scans the individual data vector one by one and decided whether the object will be gathered within the first-built leaf entries or formed a new one Kudsiati (2006).

   If a branch has already reached its capacity limit, then the node will be separated into two nodes based on the furthest object within. Then the other object will be distributed based on the closeness criterion. This process continues until the entire objects form clusters. Bacher *et al.* (2004) stated that the result of pre-clustering process depends on the order of object that has been ordered in data matrix, so that they recommend to randomize the order of the data.

2. Number of Clusters Optimization
   Bayesian Information Criterion can be used to obtain the maximum number of clusters. The Formula of BIC is given as follows:

$$BIC\,(j) = \; -2\sum_{i=1}^{j} \xi_i + m_j Log(N) \tag{1}$$

$$m_j = j\left\{2K^A + \sum_{K=1}^{K^B}(L_k - 1)\right\} \tag{2}$$

   Where:
   j : the number of clusters
   K^A : the number of numerical variable
   K^B : the number of categorical variable
   L_k : the number of category of the k^{th} categorical variable
   N : the number of observations
   The maximum number of clusters can be decided by looking at the first ratio of BIC value alteration which $<$ c1 (value of c1=0.04 based on SPSS's simulation study Technical Report (2001)).

   The opti mal number of clusters can be obtained by counting the ratio of distance measure with the formula as follows:

$$R\,(j) = \; \frac{k_{j-1}}{k_j} \tag{3}$$

$$k_j = l_{j-1} - l_j \tag{4}$$

$$l_v = \frac{(m_v \log(N) - BIC(v))}{2}; v = j, j-1 \tag{5}$$

Where,
k_j : distance when j cluster are grouped into j-1 clusters
The optimal number of clusters can be obtained by examining the greatest change of distance ratio (R(j1)/R(j2)) with R(j1) as the greatest distance change and R(j2) is the second greatest distance change. If the comparison is greater than c2 (c2= 1.15 based on SPSS's simulation study Technical Report (2001)) then, the optimal number of clusters are j2, else the optimal number of clusters are $max(j1, j2)$.

2.2 Log-likelihood Distance

Chiu *et al.* (2001) used Log-likelihood to measure the object similarity because its capability to measure the similarity of mixed categorical and numerical variable. The formula of log-likelihood distance can be seen as follows:

$$d(j,s) = \xi_j + \xi_s - \xi_{<j,s>} \tag{6}$$

$$\xi_v = -N_v \begin{bmatrix} \sum_{k=1}^{K^A} \frac{\log(\widehat{\sigma_k^2} + \widehat{\sigma_{vk}^2})}{2} - \\ \sum_{k=1}^{K^B} \sum_{l=1}^{L_k} \frac{N_{vkl}}{N_v} \log\left(\frac{N_{vkl}}{N_v}\right) \end{bmatrix}; \begin{matrix} v = j, s, \\ < j, s > \end{matrix} \tag{7}$$

Where:
$d(j,s)$: distance between cluster j and s
ξ_v : log-likelihood for cluster v
$< j, s >$: marginal index for cluster j and s
N_v : number of object in cluster v
$\widehat{\sigma_k^2}$: variance predictor of k^{th} numerical variable for all objects.
$\widehat{\sigma_{vk}^2}$: variance predictor of k^{th} numerical variable in cluster v.
N_{vkl} : number of objects in clusters v for k^{th} numerical variable and l^{th} categorical variable.

2.3 Missing value imputations

Data that has been obtained from Madiun Districts BPS still contains missing values in some Subdistricts. The value is missing in the level of villages but it is being provided in the level of subdistricts. It needs to be carefully handled if the variable is an important predictors to build the clusters. To know whether the variable is important or not we can use all villages with the complete data and cluster it first. When the output is shown we can see which variable is important to classify the villages into clusters.

There are two methods of imputations that being used in this research depends on the variable types. If the variable is not an important one we can use the common method of imputations with means for continue variables or modes for nominal variables. The means is calculated from the total value of the variables divided by the number of villages in those subdistrict.

$$\widehat{\mu_{ij}} = \frac{\sum a_i}{n_i} \tag{8}$$

Where:
$\widehat{\mu_{ij}}$: estimations value for j-th village in i-th subdistricts

$\sum a_i$: total of variable in subdistricts i
n_i : total of villages in subdistricts i
If the missing value of the data is classified as nominal variable, then it can be assumed that the variable is not exist in those specific villages.

If the missing value of the data is classified as nominal variable, then we can impute it with its modes of its sub-district.

$$\widehat{x} = \begin{cases} 0 & \text{, the variable is not existed in the village} \\ 1 & \text{, the variable is existed in the village} \end{cases} \tag{9}$$

2.4 Silhouette Coefficients

Silhouette coefficients can be used to examine the object into each clusters. This coefficient value ranged from -1 to 1. The negative value of Silhouette Coefficient means that the object measured is more related with the object in the other clusters. The lower the value of SC the further the distance between the object and the clusters.

The average of silhouette Coefficients can measure the quality of clusterization. The value also ranged from -1 to 1. If the value of Silhouette Coefficients is between -1 to 0.2 then the quality were being classified as Poor, 0.2 to 0.5 as Fair, and 0.5 to 1 as Good. The formula of Silhouette Coefficients can be seen as follows:

$$S_i = \frac{(b_i - a_i)}{\max(a_i, b_i)} \tag{10}$$

$$\bar{S} = \frac{1}{N} \sum_{i=1}^{N} S_i \tag{11}$$

Where:
S_i : Silhouette Coefficients for i^{th} object
b_i : average of minimum distance between i^{th} object in a different cluster.
a_i : average of minimum distance between i^{th} object in the same cluster.
$\bar{S}$: Average value for the Silhouette Coefficients
N : the total number of observations.

2.5 Data

This research conducted a data analytics to cluster the entire village of Madiun Districts based on their village potential. The data has been obtained from 15 "*Kecamatan dalam Angka*" of Madiun District's in 2017 from Central Bureau of Statistics ([BPS] (2017)) publication. The data describes the villages potential data for Madiun distrists in each subdistricts. The data consists of 206 observations and Methodology that being used in this research is Two Step Cluster with 15 variables consists of 9 numerical varibles and 6 categorical variables shown in Table 1.

2.6 Analysis Stage Procedure

In this research, the analysis has been conducted with IBM SPSS 16 software. The analysis step were described as follows:

1. Describe the general characteristics of Madiun District.

Table 1: VARIABLES USED IN THE ANALYSIS

Variable	Name of Variable	Scale
X1	Village status (village/Formal village)	Categorical
X2	Village distance to Sub-districts capital	Numerical
X3	Village Population at the end of 2016	Numerical
X4	Number of family	Numerical
X5	Number of elementary school (SD/MI)	Numerical
X6	Number of public health facility (*poskesdes,polindes,posyandu*)	Numerical
X7	Number of worshiping facility	Numerical
X8	Number of Couples of Childbearing Age (PUS)	Numerical
X9	Total farming area	Numerical
X10	Art Club Existence	Categorical
X11	Market existence	Categorical
X12	Type of the longest road surface	Categorical
X13	Rural banks existence	Categorical
X14	Cooperation existence	Categorical

2. Analysis were being run with only villages that has complete variables. Villages with missing value needed to be excluded in this examination. The number of villages that has been run in this step is 136 villages with complete variables, the rest were excluded.

3. After that, Chi-Square and T-test value for each variables were examined. Variables were classified into the important and not important one based on it.

4. Imputed the missing data with means and modes.

5. Chi-Square and T-test value of each variables were examined to determine the important variables.

6. Characteristics of each cluster built were described.

3 Results and Discussions

3.1 Exploration

Madiun District has 206 Villages which consists of 198 villages and 8 formal villages. Two step clustering method has been used to cluster the villages in Madiun Districts with Bayesian Information Criterion (BIC) value to determine the optimal number of built cluster. In this case, there are no different outcome when the evaluator for the optimal number of built cluster is AIC or BIC, so then BIC has been chosen subjectively to evaluate the optimal number of built Cluster.

Distance measurement that being used in this research is log-likelihood distance. Log-Likelihood distance can overcome the effect of mixed variables used (Categorical and Numerical variables) rather than Euclidean distance.

In this data still contain missing value for certain variables, such as total farming area and number of worshipping place as continues variables also art club existence as the nominal variables. In order to deal with the missing value, the analysis need to be ran with only villages that has complete variables. Villages with missing value needs to be excluded in this examination. The number of villages that being ran in this step is 136 villages, the rest are being excluded.

Table III shows that the maximum number of clusters that can be built in the first step is 5. It can be referred to the value of Ratio of BIC Changes that less than 0,04 is in the number of clusters 5. In the second step, we can examine the most significant value for Ratio of Distance Measures is coming from the number of clusters 4 and 8 with the value 1.59 (R4) and 1.43 (R8). The ratio of R4/R8 is 1.118 less than c2=1.15 so that the optimal number of built cluster is 8, but it has been limited by 5 as the maximum number of built cluster. So, it can be concluded that the optimum number of built cluster is 4.

Table 2: RATIO OF BIC CHANGE VALUE

Number of Clusters	Ratio of BIC Changes	Ratio of Distance Measures
1		
2	1.000	1.134
3	.795	1.255
4	.484	1.597
5	.028	1.082
6	-.030	1.287
7	-.188	1.001
8	-.188	1.428
9	-.352	1.053
10	-.372	1.140

3.2 Characteristics of the Clusters

The number of clusters built is 4 with the Cluster distribution can be seen in the Table IV, where cluster 1 consists of 17 villages, cluster 2 consists of 48 villages, cluster 3 consists of 61 villages, and cluster 4 with 80 villages.

Based on Table 3, number of Villages in cluster 1 is 17. The important variables for Cluster 1 consist of 5 variables (Figure 2), which are existence of art club, number of child-bearing age couples, number of Elementary School, number of Family, and Village population. Cluster 1 is dominated by highest mean in some variables. Those variables are village population, number of family, number of elementary school, number of public health facility, number of child-bearing age couples and total of farming area. This cluster also has more existence upon not existence in the market, Rural bank, and cooperation and type of longest road.

As we can see in Table III, Cluster 1 has the lowest number of member compare to the other clusters. Cluster 1 has the greatest amount of human resources in between the other clusters. Those amount can be shown by the number of Village Population, Family, and number of child-bearing age couples with means of 5950.8, 2013.5, and 1473,8 respectively.

Cluster 1 also has greatest value in mean on the public facility such as number of elementary school and health facility with the value of 4.4706 and 7.588 respectively. farming area in this cluster also classified as the highest area among the other with means of 769.56 ha. Variable of Art club existence is included in the significant variable, but there is no single villages in cluster 2 that have an Art club.

Variable in cluster 1, in general, is in the highest side among the others, which means the village member of this cluster is well developed. This clusters have a great amount of populations but they also have sufficient facility to support the community. This cluster member can be considered as a leading example for the other village in the other cluster.

Cluster 2 consist of 48 villages with 5 significant characteristics as can be seen in Figure 3. The significant variable of cluster 2 is the existence of art club, the number of child-bearing age couples, the number of Elementary school, the number of village population, and the number of family. in the contrary with cluster 1, cluster 2 has the lowest means in some variables. Those variables are village population, number of family, elementary school, worshipping place, and child-bearing age couples.

Cluster 2 has the lowest human resources from the other clusters, those has been shown by the value of village popultion, number of family, number of child-bearing age couples with the value of 2505, 866.7, and 527.06 respectively. Number of Elementary school and worshipping place also shown the lowest value among the others with means 1.583, and 12.5208 repectively. Distance of cluster 2 member to sub districts is the greatest among the others with means 8.0417. it means that majority of the villages are located far enough from the sub district capital.

In General, Cluster 2 member can be classified as low-developed villages group. Since they have fine lowest amount of facility and human resources. This cluster member can be a perfect candidate to be developed both on the facility and human resources.

Cluster 3 consists of 61 villages. Based on t-student test, the variables that considered to be significant are number of public health facility, total of farming area, and the number of elementary school. Based on Chi-Square test, the existence of art club, rural banks, type of longest road, and existence of cooperation are considered as the important variable for this cluster. As can be seen in Table III, two of the numerical variables of this cluster are having the least means between the other. Those variables are number of public health facility and total farming area with the value of the means are 4.7377 and 185.68, which support that those two variables become the important varibale in this cluster.

Means of the village population is 3139.2 people per village. Means of the number of Family variable is 1073.5 family per village. Means of Couples of Child Bearing Age is only 639.6 per village. Means for total of Worshipping place is 15.8525 facility per

Table 3: CLUSTER DISTRIBUTION

Description	Cluster			
	1	2	3	4
Number of Status				
Villages	17	48	57	76
Adm. Villages	0	0	4	4
Existence of Art Club				
Not Exist	17	19	51	0
Exist	0	29	10	80
Existence of Market				
Not Exist	8	42	45	52
Exist	9	6	16	28
Type of Longest Road				
Not Asphalted	4	27	0	4
Asphalted	13	21	61	76
Existence of Rural bank				
Not Exist	3	0	17	0
Exist	14	48	44	80
Existence Cooperation				
Not Exist	5	29	0	1
Exist	12	19	61	79
Means				
Distance to subdistrict	7.3529	8.0417^a	5.1967	3.9375^b
Village Population	5950.8^a	2505^b	3139.2	3910.1
Number of Family	2013.5^a	866.77^b	1073.5	1349.4
Number of Elementary School	4.4706^a	1.5833^b	1.7541	2.0500
Number of Public Health Facility	7.5882^a	5.1042	4.7377^b	6.6375
Number of Worshiping place	20.1176	12.5208^b	15.8525	21.1625^a
Number of Child-Bearing Couples	1473.8^a	527.06^b	639.64	678.80
Total of Farming Area	769.56^a	205.9	185.68^b	243.85

^{a}Lowest Mean, ^{b}Biggest Mean

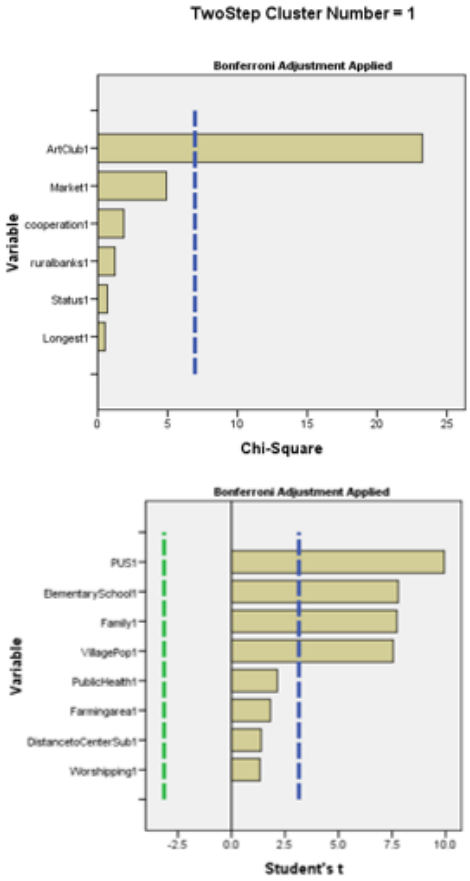

Figure 1: Predictor Importance of Cluster 1

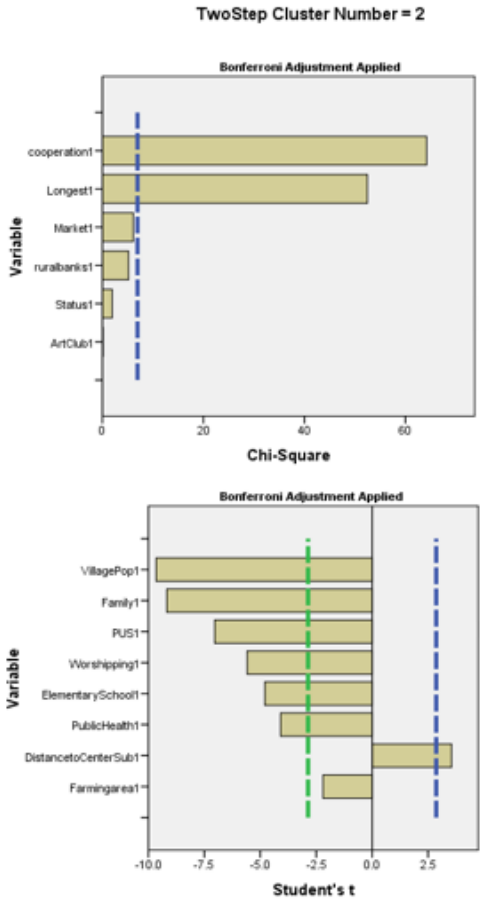

Figure 2: Predictor Importance of Cluster 2

village, means of Total of Public health 4.737 facility per village. Means of the number of Elementary School is 1.3882 school per village. This number is small since there are 2 villages that has no elementary school building. Those villages are Kuncen and Pucanganom. This facility is crucial for every village to have since it may affect the activity for each village to get an education. If the education facility is not available in their own village, their children have to travel to the nearest neighboring villages to get education. It will become another obstacle to make the village society get a proper education.

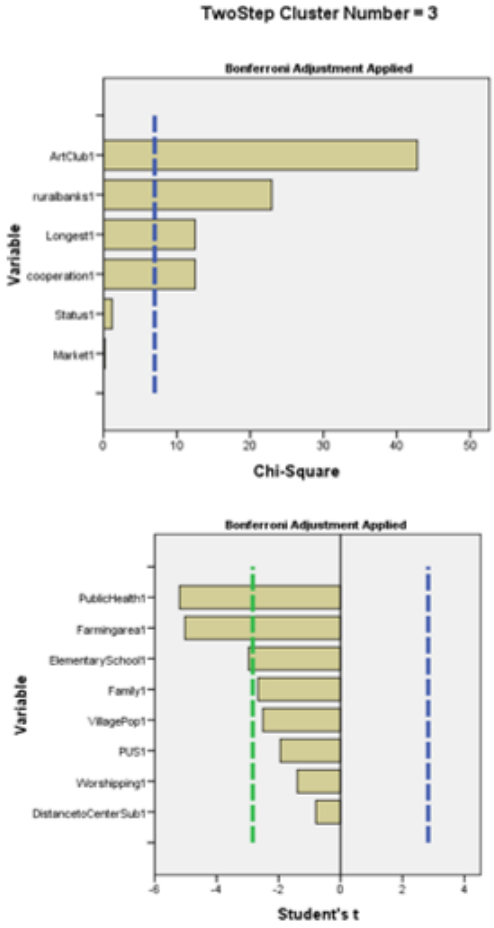

Figure 3: Predictor Importance of Cluster 3

Compared to the other clusters built, Cluster 3 characteristics are nearly similar with cluster number 2. Cluster number three also can be considered to be cluster that needed attention t be developed.

As shown in Figure 4 that the important variables in cluster 4 is the existence of artclub, cooperation, rural bank, distance to the center of subdistrict, number of Public health, and worshipping places. This information also being supported by information in Table 3. The mean distance to the center of subdistrict in this cluster is the lowest between the others with the value 3.937 km per village. Which means the member of this cluster are villages that located near the center of subdistricts. The resident in this cluster have previleges for easiness access to the governmental facility.

The number of worshipping places is also showing the highest value among clusters with value 21.1625 and the number of public health facility is also quite high in the value

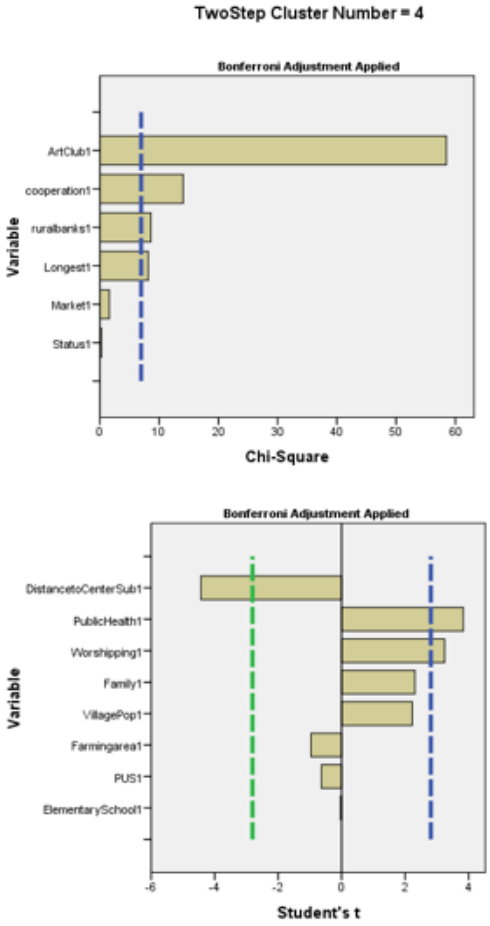

Figure 4: Predictor Importance of Cluster 4

6.637 facility per villages. With those characteristics cluster 4 can be considered as the cluster that need a little attention to be developed.

3.3 Cluster Quality

Silhoutte coefficients has been used to determine the quality of built clusters. In Figure 5, can be seen that the quality is in fair Figure 5 Silhouette Coefficient category, with the value of silhouette coefficient is 0.3. With this value, it can be interpreted that the member of each clusters are fairly closed to the center of respected clusters.

Figure 5: Silhouette

4 Conclusion

Cluster analysis for villages in Madiun District generate 4 built clusters. Cluster that need to be developed for facility existency can be focused on the member of Cluster 2. Cluster 2 need to have more attention to be developed because they are having the least means for the majority of its variable. The development on facility also can be focused on the village that does not have a specific facility at all such as elementary education facility. Those cases can be found in Pucanganom village of Kebonsari subdistrict and Kuncen village of Mejayan Sub-district. In the other hand, cluster 1 having a massive amount on human resources. It has the biggest means for human resources compared to the other cluster and they also have biggest means in the number facility. Instead of developing facility, the proper development that can be done for member of cluster 1 is human resources development which include organizing and managing the community, so that the community can get an advantage from the amount of those resource.

References

Alam, Y. & Ambarwati, A. N. (2017). *Analisis Cluster Pada Desa/Kelurahan Di Kabupaten Wonosobo Berdasarkan Data Potensi Desa Tahun 2015*. Semarang (ID): AIS Muhammadiyah Semarang.

Bacher, J., Wenzig, K., & M, V. (2004). *SPSS TwoStep Cluster: A First Evaluation*. Germany: Friedrich Alexander-Universität Erlangen-Nurnberg.

[BPS] (2017). *Kabupaten Madiun dalam Angka 2016*. Jakarta (ID): Badan Pusat Statistik.

Chiu, T., Fang, D., Chen, J., Wang, Y., & Jeris, C. (2001). A robust and scalable clustering algorithm for mixed type attributes in large database environment. In *Proceedings of the seventh ACM SIGKDD international conference on knowledge discovery and data mining*, pages 263–268.

Kudsiati (2006). *Pengkajian Keakuratan TwoStep Cluster dalam Menentukan Banyaknya Gerombol Populasi [tesis].* Bogor (ID): Institut Pertanian Bogor.

Mattjik, A. A. & Sumertajaya, I. (2011). *Sidik Peubah Ganda dengan Menggunakan SAS.* Bogor (ID): IPB Press.

Putri, W. D. Y. (2005). *Penerapan Metode Two Step Cluster dalam Analisis Gerombol (Studi kasus: data Potensi Desa Sensus Ekonomi 2003 wilayah Jawa Barat) [skripsi].* Bogor (ID): Institut Pertanian Bogor.