The Clustering of Households in Madura Based on Factors Affecting Their Ingestion of Clean Water Using *Similarity Weight and Filter Method*

Astarani Wili Martha and Ismaini Zain
Statistics Departement, Faculty of Mathematics, Computing and Data Science,
Institut Teknologi Sepuluh Nopember (ITS)
Jl. Arief Rahman Hakim, 60111 Surabaya Indonesia
e-mail: *ismaini_z@statistika.its.ac.id*

Abstract— Clean Water and Sanitation is one of SDGs' indicators that relates to human' demand for clean water. Three of four regencies in Madura Island reportedly have suffered in drought, thus it leads this research to fulfill Madura people need of water. Madura Island has 3097 households in need of water. However, not all households could fetch their need. This research aims to classify the households of Madura Island regarding factors which affect their ingestion of clean water using cluster analysis. There are clustering numerical data and categorical data. Therefore, this research uses Similarity Weight and Filter Method. SWFM is one of clustering mix methods in which there are clustering numerical, using hierarchical ward, and clustering categorical, using k-modes. To analyze the clustering numerical data, there are 3 variables and it gains two optimum groups by using ward method with pseudo-F 1001,172. Clustering categorical analysis uses 6 variables with k-modes and gains three groups and SWFM gains five groups. Five groups are selected because they produced the smallest ratio 0,006627 in the group.

Keywords: Clean Water, Cluster Analysis, Households, K-modes, SWFM.

I. INTRODUCTION

Sustainable Development Goals (SDGs) is an advanced developing program which has 17 aims and 169 targets of development. The aim of SDGs is for the government to overcome the problem of underdevelopment of country in the world, whether it is a developed or developing country, regarding the world, changes issues such as environmental damage, climate change, social protection, deflation of natural resources and other development related to poverty. In the 17 aims, there is one that aims for clear water demand called *Clean Water and Sanitation*. The issue regarding clear water demand must be resolved in order to fulfill SDGs in 2030.

Economic development cannot be separated from the study of access to clean water. Development is the government's capability to fulfill basic needs, such as food, shelter, health, and protection (including access to clean water) [1]. If one of those does not fulfill, people will experience the condition of poverty. According to UNDP, poverty is a situation or condition in which a person (individual) does not have income in order to fulfill basic needs and it consists of the right to enjoy a dignified life and right recognized in the laws [2]. One of the basic needs is access to clean water. Suparman [3] stated that people that have the status of poor households are even more difficult to get clean water although it is the poor households who get more bills than non-poor households in order to get clean water.

According to *Badan Penanggulangan Bencana Daerah* (BPBD), or Council of Regional Disaster Management, there are 15 regencies in East Java that suffered in drought including 3 regencies in Madura Island. This issue becomes one of the problems in the inequality of clean water distribution in each region, notably the area on Madura. Based on the 2017 PDAM's (local water company) performance, 3 of 4 regencies in Madura are less healthy than others [4]. Indeed, it is a challenge for government and PDAM to improve access to clean water service, as it is stated in the 2019 RPJMN (national mid-term development program) that the government aims to give a 100% access of clean water for its people.

In the representation of access to clean water in Madura, clustering into groups is needed in order to analyze demands of clean water based on factors which affect their ingestion of clean water. This research uses numerical and categorical research variables. *Similarity Weight and Filter Method* is a method of clustered mix dataset consists of numerical and categorical.

Based on the description above, this research would be conducted with clustering factors which affect the households in Madura regards to their need for clean water using *Similarity Weight and Filter Method* (SWFM). This research is expected to be a suggestion for the government in handling the needs of clean water based on related factors in the households of Madura Island.

II. LITERATURE REVIEW

A. Descriptive Statistics

Descriptive Statistics is a method relates to the collection and representation of a data group in order to provide information. Descriptive statistics only provide information regarding available data and cannot be used to generalize the main data group. Descriptive statistics also presented in the form of a table, diagram, graphic and other quantities [5]. In this method, the data would be divided into two which are categorical and numerical.

B. Clustering Analysis

Cluster Analysis is one of the multivariate analyses which aim to place a set of objects into twoor more clusters based on their similarity characteristics [6]. The purpose of this analysis is to cluster the objects of observation into several clusters based on their characteristics. Cluster analysis classifies objects and makes each object impending to others with similar characteristics [7]. The result of cluster analysis is influenced by several things, including clustered objects, observed variables, similarity and inaccuracy measures, and clustering methods.

The similarity and dissimilarity are generally measured by distance. One of the factors that greatly influence the result of the clustered group is the distance between object of observation [8]. Here is the following method of distance measurement between *i* object $\left(x_i\right)$ with *j* object $\left(x_j\right)$ based on clustering variable characteristic.

a. *Euclidean Distance*

Clustering numerical data based on a measure of dissimilarity or distance. The measure of dissimilarity that commonly used is *Euclidean* Distance. The use of *Euclidean* distance is relatively easy to understand and can be used on data that has more than two variables as it shown in the following equation (1).

$$d_{ij} = \sqrt{\left(x_{i1}-x_{j1}\right)^2 \left(x_{i2}-x_{j2}\right)^2 + \ldots + \left(x_{im}-x_{jm}\right)^2} \quad (1)$$

Where,

i = $\left(1,2,\ldots,n\right)$ and j = $\left(1,2,\ldots,n\right)$.

C. Clustering Numerical Data

The hierarchical method is used if the number of clusters to be formed is unknown beforehand and the number of observation is not too big. There are two techniques in hierarchical cluster analysis; *divisive* and *agglomerative*. Some of the cluster techniques between cluster as stated as follows [7].

Ward method uses basic considerations to minimize lost of information from the merger of two clusters. The sum of quadratic between two clusters for all variables is the distance between two clusters thus this method minimizes variance in clusters. The value of ESS shows in equation (2).

$$ESS = \sum_{j=1}^{N} \left(x_j - \bar{x}\right)' \left(x_j - \bar{x}\right) \quad (2)$$

Determining the optimum number of clusters is an important step after the grouping process. This stage is referred to as clustering validation [8]. The R-squared index is one index that can be used to determine the optimum number of groups in hierarchical clustering [9]. The index involves calculating the diversity of data in terms of total diversity, group diversity, and diversity between groups. The validity index to determine the optimum number of groups in hierarchical clustering can be written as follows.

$$SST = \sum_{l=1}^{m_{numerical}} \sum_{i=1}^{n} \left(x_{il} - \bar{x}_l\right)^2 \quad (3)$$

$$SSW = \sum_{c=1}^{C} \sum_{l=1}^{m_{numerical}} \sum_{i=1}^{n_c} \left(x_{ilc} - \bar{x}_{lc}\right)^2 \quad (4)$$

$$SSB = SST - SSW \quad (5)$$

where,

$m_{numeric}$: Number of numerical variables in observation,

c : Number of clusters formed in observation,

n : Total number of observation objects,

n_c : Number of members in cluster c for c=1,2,…,C

$\bar{x}_l$: Average of all objects in variable *l* for $l = 1,2,\ldots,m_{numerical}$,

$\bar{x}_{lc}$: Average variable *l* in group *c* for $c = 1,2,\ldots,C$

R^2 = 0 index that is no difference between clusters, while R^2 = 1 shows a significant difference between clusters formed.

$$R^2 = \frac{SSB}{SST} = \frac{\left[SST - SSW\right]}{SST} \quad (6).$$

Determination of the number of clusters formed can be seen based on the maximum value of Pseudo-F. The formula used to calculate the value of pseudo-F statistics shows in equation (7).

$$Pseudo - F = \frac{\left(\dfrac{R^2}{c-1}\right)}{\left(\dfrac{1-R^2}{n-c}\right)} \quad (7).$$

D. Clustering Categorical Data

Clustering categorical data can be done by measuring the similarity of data using non-hierarchical k-modes. Based on Kaufman and Rousseeuw [10], suppose X and Y are two data with categorical type features. The size of the dissimilarity between X and Y can be measured by the number of incompatibility values of features that correspond from two data. The smaller the value of the incompatibility, the more similar the two data. The following is the formula used as in equation (8).

$$d\left(X,Y\right) = \sum_{j=1}^{r} \delta\left(x_j, y_j\right) \quad (8).$$

E. Clustering Mixed Data

SWFM clustering has a concept similar analysis by clustering ensemble in general, this method is an outgrowth of the clustering ensemble that has a different algorithm to get the final stages of cluster formation. The similarity weight method stage used similarity measures that include weighting factors on similarity size formulas. The weight given depends on the number of observation (n_i or n_j). The formula used to calculate the size of similarity. Between *i* object and *j* object is as in equation (9) [11].

$$sim\left(x_i, x_j\right) = \sum_{i \le n_i, j \le n_j} \frac{S_{ij}}{\max\left(n_i, n_j\right)}, \ i \ne j \quad (9)$$

$$S_{ij} = \left|\frac{X_i \cap X_j}{X_i \cup X_j}\right|$$

X_i : The *i*th observation set with $X_i = \left\{x_{1i}, x_{2i}, x_{3i}, \ldots, x_{ki}\right\}$

X_j : The *j*th observation set with $X_j = \left\{x_{1j}, x_{2j}, x_{3j}, \ldots, x_{kj}\right\}$

$\left|X\right|$: Cardinal number or number of members of the set *X*

m_k : Number of categorical variables in observation

n : Total number of observation objects

n_i : Number of members in the *i* cluster

n_j : Number of members in the *j* cluster.

the results of clustering numerical and categorical data are combined to obtain the final clustering using the filter method as in equation (10) [11].

$$F\left(X_i, X_j\right) = \sum_{i=1}^{n_i}\sum_{j=1}^{n_j} w_{ij} d\left(X_i, X_j\right) \qquad (10).$$

F. Validation Clustering

The performance of the results of clustering for variables with a numerical scale can be seen from the ratio of the value of S_W and S_B [12]. By using variable mean values, the standard deviation in clusters $\left(S_W\right)$ and standard deviation between clusters $\left(S_B\right)$ can be formulated in equations (11) and (12).

$$S_W = \left[MSW\right]^{1/2} \qquad (11)$$

$$S_B = \left[MSB\right]^{1/2} \qquad (12)$$

As with numerical data, the performance of clustering with categorical data is also based on a comparison of the ratio between standard deviation in groups $\left(S_W\right)$ and standard deviation between clusters$\left(S_B\right)$. Where if the comparison ratio gets smaller, the performance of categorical data is getting better because of maximum homogeneity in clusters and maximum heterogeneity in clusters.

G. Normality Test

This normality test uses *Kolmogorov Smirnov* and *mshapiro* test.

Kolmogorov Smirnov test is used to check the distribution of a variable. The *mshapiro* test method used to test whether the data is the multivariate normal distribution.

Hypothesis

H_0 : Data follow a normal distribution

H_1 : Data do not follow a normal distribution.

The *Kolmogorov Smirnov* test statistics:

$$D = \underset{x}{Sup}\left|F_n(x) - F_0(x)\right| \qquad (13)$$

Where,

$F_n\left(x\right)$: sample cumulative distribution

$F_0\left(x\right)$: empirical distribution (cumulative distribution under H_0 $P(Z<Z_i)$)

Critical Region : Reject H_0 if $D > D_\alpha$

The *mshapiro test* in equation (14) is as follows.

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} \qquad (14)$$

Where, $\left(a_1, a_2, \ldots, a_n\right) = \dfrac{m^T V^{-1}}{\left(m^T V^{-1} V^{-1} m\right)^{1/2}}$.

The decision to reject H_0 is taken with the determination of a significant level of α then H_0 is rejected if *W* is less than *a* [7].

H. Kruskal Wallis Test

The *Kruskal Wallis* test is a nonparametric technique that is used to determine whether the sample is from an identical population. Some assumptions for this test are independent observations and at least ordinal scale of measurement. The hypothesis used is.

H_0 : the samples (clusters) are from identical populations

H_1 : at least one of the samples (clusters) comes from a different population than the others.

The Kruskal Wallis test statistics in equation (15).

$$H = \frac{12}{N\left(N+1\right)}\sum_{i=1}^{g}\frac{R_i^2}{n_i} - 3\left(N+1\right) \qquad (15)$$

Where,

R_i : Sum of ranks in the i^{th} cluster

n_i : number of observation in cluster *i*

N : the total number of observation across all clusters

K : number of clusters

Decision making uses the value of *H* compared to the value of $\chi^2_{(0,05;K-1)}$. The decision rejects H_0 if the value of *H* is greater than $\chi^2_{(0,05;K-1)}$ or the value of $\chi^2_{statistic}$ is greater than $\chi^2_{(0,05;K-1)}$ [13].

I. Factors Affecting Their Ingestion of Clean Water

According to Winarna [13], in her research stated that the factors affecting their ingestion of clean water of household customers in Karanganyar Regency are family income, total household expenditure, household members, education of the family head and the presence or absence of PDAM water sources. Expenditures for water consumption do not differ greatly from other expenses in the household. Research conducted by Joachim and Hillenbrand [15], household water consumption is affected by water rates or prices, per capita income, a number of household heads, population, season or climate, and temperature. In addition, according to Darr and Kamen [16], several factors that were considered in influencing household water consumption included household size, per capita income, area development, meter type, respondent education, space density and service coverage. But according to the Ministry of Women, this is not the case. Often women have a lot to do with water needs both for the benefit of individuals and households.

III. Case Study

The data used are secondary data. Data were obtained from SUSENAS (National Socio-Economic Survey) East Java in 2016. The research units used were households in four Regencies on the island of Madura with a sampling of 3,097 households. The variable observation used were 9 variables including three numerical scales and six categorical scales.

Table 1 The Variable Observation

Variable	Name	
X_1	Total household expenditure	-
X_2	The proportion of female household members	-
X_3	Number of household members	-
X_4	Education degree of the head of household	1 : Not Graduated from Elementary School (NGES) 2 : Low 3 : Middle 4 : High
X_5	Drinking water sources	1 : Can be purchased (CP) 2 : Cannot be purchased (CNP)
X_6	The water source for cooking	1 : Can be purchased (CP) 2 : Cannot be purchased (CNP)
X_7	The water source for	1 : Can be purchased (CP)

Variable	Name	
	bathing/washing	2 : Cannot be purchased (CNP)
X_8	Water supply system	1 : Use (U) 2 : Not Use (NU)
X_9	Water bill payment for PAM	1 : Not Pay (NP) 2 : Pay (P)

The operational definition of the variable observations are as follows:

1. Total expenditure per month is the amount of expenditure spent by households for consumption of other needs (Million).
2. The proportion of female household members is the number of female household members who live in the residence (Percent).
3. The number of household members is the number of household members living in the residence (People).
4. Education degree of the head of household is the highest education completed by the head of the household.
   Category:
   1) NGES (Not Graduated Elementary School) is not complete or graduated from elementary school.
   2) Low is the education level in elementary school/equivalent graduate and middle school/equivalent graduate.
   3) Middle is the education level in high school/equivalent graduate.
   4) High is the education level in college student.
5. Drinking water source, the Water source for cooking, the Water source for bathing/washing are the main water source used by households on Madura.
   Category:
   1) CP (Can be Purchased) is water sources that can be purchased such as branded water, refill water, plumbing meters and retail plumbing.
   2) CNP (Cannot be Purchased) is water sources that cannot be purchased such as borehole water/pumps, protected wells, unprotected wells, protected springs, unprotected springs, rainwater, surface water, and others.
6. The water supply system is a system of components which provide water.
   Category:
   1) U (Use) is households using a piped water system or public hydrant.
   2) NU (Not Use) is households do not use or do not know the piped water system / public hydrant.
7. Water bill payment is a household that pays for PAM (local water company).
   Category:
   1) NP (Not Pay) is households that do not use and do not pay PAM.
   2) P (Pay) is households that use and pay for PAM.

The analysis steps in this research are as follows:

1. Make descriptive statistics for numerical data variables by determining the maximum and minimum value, variance, median and mean of each variable. Whereas for categorical data with graphs.
2. Clustering households on Madura using the Similarity Weight and Filter Method.
   a. Divide the variables observation into categorical and numerical data.
   b. Cluster numerical data using the agglomerative ward hierarchical method. The distance used is Euclidean distance.
   c. Determine the optimum number of clusters, calculate and select clustering performance based on the maximum pseudo-F value for each clustering numerical result.
   d. Cluster variables are categorical data using the *K-modes* method.
   e. Combining cluster results in stages c and d.
   f. Processing data using *Similarity Weight* Method.
   g. Clustering data using *Similarity Analysis*.
   h. Determine the *Filter Algorithm* for categorical and numeric data.
   i. Clustering data using *Similarity Weight and Filter Method.*
   j. Determine the optimum number of clusters based on the lowest *Sw* and *Sb* ratio.
3. Testing data with *normality test* and *kruskal wallis test.*
4. Interpret the result and make a conclusion.

IV. Results and Discussion

A. Characteristics of Numerical Data

In this research, numerical variables are used based on factors household water demand including total expenditure (X_1), the proportion of female household members (X_2), and the number of household members (X_3). The results of descriptive statistical analysis are shown in Table 2.

Table 2 Characteristics Data for Predictor Variables

Variable	Mean	Variance	Median	Min	Max
X_1 (million)	2.414	2.068	1.929	0.179	29.612
X_2 (percent)	0,55	0,22	0,5	0	1
X_3 (people)	4	2	4	1	14

The total expenditure variable has an average of 2,414 million rupiahs per household with a total expenditure of at least 179 thousand rupiahs with households living in Bangkalan Regency. While the largest total expenditure amounted to 29,612 rupiahs with households in Sumenep Regency.

The proportion of female household members has an average of 0.55 percent with the smallest proportion of 0 or no household member living. While the largest proportion of 1 percent with the highest proportion of female household members is Sumenep Regency.

The number of household members has an average of 4 people with a minimum number of household members of 1 person, the largest number of members of a house in Sumenep Regency. While the largest number of household members is 14 people, with households living in Pamekasan Regency.

B. Characteristics of Categorical Data

In this research, categorical variables were used education degree of the head of household, drinking water source, the water source for cooking, the water source for bathing/washing, water supply system, and water bill payment for PAM.

The categorical data in the education degree of the head of household was low education at 52.57%. Whereas water sources such as drinking water source, the water source for cooking and bathing or washing use more water sources that cannot be purchased. The water supply system is more commonly used is piped water system/public hydrant and the majority of households not paying PAM.

C. Clustering Numerical Data

Clustering for numerical data uses the average linkage technique with each observation object. The optimum cluster number is in 5 clusters with the ward method. The number of clusters formed between two to six clusters was carried out using software R. After the results of grouping from two to six groups were obtained, *the Pseudo F-statistics* values were obtained. The results of the *pseudo-F* value obtained by the optimum cluster number formed 2 clusters with the largest pseudo-F value is 1001,172, because the pseudo-F value is greater than the number of other clusters.

Table 3 Characteristics Data for Clustering Numerical Data

Variable	Cluster 1	Cluster 2
X_1 (Million)	3.36	1.13
X_2 (Percent)	0,51	0,61
X_3 (People)	4	3

The division of the number of cluster members with the ward hierarchy method for the number of 2 clusters is 1,782 and 1,315 households.

While the characteristics used are the average values of each cluster formed. The following results from the average number of cluster 2 are as follows.

Table 3 shows the average values of each cluster member. Based on Table 3, cluster 1 looks to have the highest average value in the variable total household expenditure and the number of household members. Then for cluster 2, it has the smallest average value for the variable total household expenditure and the number of household members. But the average value for the variable proportion of female household members for cluster 2 is greater. Characteristic results from clustering numerical scale variable, it can be seen from each cluster showing the characteristic results based on household economics on Madura. This can be seen from the characteristics of households in cluster 2 which are dominantly included in the low economy. Then for the dominant cluster 1 characteristics, including the high economy.

In addition to the characteristics of each cluster per household on Madura, it is also necessary to know the results of each cluster of households living in the regency on Madura based on the criteria or characteristics based on the characteristics formed. Results cluster of households with a more low economy are households living in Sumenep Regency.

D. *Clustering Categorical Data*

Clustering on categorical variable data by using the *k-modes* method with each group of observation objects as a single group with a single member. This study uses R software to analyze *k-modes*. The value of *k* or the number of groups used in this analysis is determined by 3. The clustering of more data is included in the number of clusters 1 so that each division of the number 3 clusters are 1.975, 723, and 399 households respectively.

The result of clustering based on characteristics with the percentage of the most dominant categories is as follows.

Tabel 4 is the percentage presentation of the most dominant category of variables in the cluster. Cluster 1 can be seen that most of them are not graduated elementary school with the use of drinking water, cooking water and bathing/washing water, the most of which is water that cannot be purchased then the water supply system used is piped system or public hydrants and for households not making payments PAM water. Cluster 2 can be seen mostly with a low education degree with sources of drinking water, cooking water, and bathing/washing water is the most widely used is water that cannot be purchased, then the water supply system does not use piped water systems or public hydrants and for households not to pay for PAM. Cluster 3 is mostly low-educated degree then with the sources of drinking water, cooking water and bathing/washing water that is used is water that can be purchased, then the water supply system uses piped system or public hydrants and for households to make PAM payments.

Table 4 Characteristics Data for C lustering Categorical Data

Variable		Cluster 1	Cluster 2	Cluster 3
X_4	NGES	**48,81%**	0,00%	16,29%
	Low	39,59%	**86,86%**	**54,64%**
	Middle	8,15%	11,34%	12,53%
	High	3,44%	1,80%	16,54%
X_5	CP	9,47%	7,19%	**97,99%**
	CNP	**90,53%**	**92,81%**	2,01%
X_6	CP	0,96%	0,69%	**96,24%**
	CNP	**99,04%**	**99,31%**	3,76%
X_7	CP	0,15%	0,14%	**79,20%**
	CNP	**98,85%**	**99,86%**	20,80%
X_8	U	**70,78%**	0,00%	**90,48%**
	NU	29,22%	**100,00%**	9,52%
X_9	NP	**87,80%**	**89,21%**	15,29%
	P	12,20%	10,79%	**84,71%**

According to characteristic clustering categorical data, it was found that the criteria or characteristics of each group in a row were the low economies, medium economy, and high economy. The cluster results of more low-income households are households that live in Pamekasan Regency.

E. *Clustering Mixed Data*

Clustering mixed data uses SWFM (Similarity Weight and Filter Method) by clustering each of the results of the previous clustering (numerical data and categorical data). The first step carried out in the analysis of the SWFM ensemble clustering for mixed data is by clustering each type of data using their respective methods.

The best grouping results obtained the smallest ratio value in the number of groups of five amounting to 0.006672. The division of the number of cluster members with the SWFM method for the number of 5 clusters is 1,063; 1,395; 276; 272; and 127 households.

Characteristics of numerical and categorical variable observations from members resulting from grouping on SWFM analysis in each cluster are presented in Table 5 and Table 6 as follows.

Table 5 Characteristics of Numerical Data in The Clustering SWFM Result

Variable	**Characteristics Cluster**				
	1	**2**	**3**	**4**	**5**
X_1 (Million)	3.30	1.71	1.20	4.41	1.23
X_2 (Percent)	0,51	0,58	0,57	0,51	0,64
X_3 (People)	4	3	3	4	3

Table 5 shows the average values of each cluster member. Based on Table 5, cluster 4 and cluster 3 have the highest average value in the variable total household expenditure and number of household members. Then for cluster 3 and cluster 5, it appears to have the smallest average value for the variable total household expenditure and the number of household members. However, the average value for the variable proportion of female household members for cluster 5 and cluster 2 is greater than the other clusters.

Table VI is the percentage presentation of the most dominant categories of variables in the cluster results. From the results of the cluster, it was found that the dominant education degree of the head of household was low education. Cluster 1, cluster 2 and cluster 3 can be seen that most of the sources of drinking water, cooking water and bathing/washing water are water that cannot be purchased and most of them do not pay PAM. Cluster 2 and cluster 3 mostly do not use a water supply system. Cluster 4 and cluster 5 mostly use water that can be purchased for source of drinking water, cooking water and bathing/washing water and most households pay PAM. Households in cluster 1, cluster 4 and cluster 5 tend to use a water supply system.

Table 6 Characteristics of Categorical Data in The Clustering SWFM Result

Variable	Characteristics Cluster				
	1	2	3	4	5
X_4	Low (44,03%)	Low (50,99%)	Low (90,22%)	Low (52,21%)	Low (50,84%)
X_5	CNP (84,01%)	CNP (95,66%)	CNP (96,38%)	CP (98,53%)	CP (96,85%)
X_6	CNP (98,40%)	CNP (99,48%)	CNP (100%)	CP (96,69%)	CP (95,28%)
X_7	CNP (99,91%)	CNP (99,85%)	CNP (99,64%)	CP (84,19%)	CP (68,50%)
X_8	U (76,39%)	NU (56,88%)	NU (100%)	U (94,12%)	U (82,68%)
X_9	NP (87,02%)	NP (88,67%)	NP (90,22%)	P (79,02%)	P (96,85%)

By knowing the characteristic results of grouping using the SWFM method, it can be seen from each cluster showing the characteristic results based on the household economy in Madura. This can be seen from household characteristics, namely

Cluster 1: High economy.
Cluster 2: Medium economy.
Cluster 3: The lowest economy.
Cluster 4: The highest economy.
Cluster 5: Low economy.

In addition to the characteristics of clusters per household in Madura, each criterion or characteristic is obtained based on the economy of households living in the Regency on Madura. Following are the results of household clusters living in regencies on Madura for clustering mixed data can be shown in Table 7.

Table 7 The Result of Clustering Mixed Data in Households in Every Regency on Madura

Regency	Cluster					Total
	1	2	3	4	5	
Bangkalan	307	291	70	61	17	746
Sampang	300	292	64	80	22	758
Pamekasan	255	355	44	56	47	757
Sumenep	201	421	98	75	41	836
Total	1.063	1.359	276	272	127	3.097

Table 7 is the presentation of the results of household clusters living in each regency using the SWFM method based on the household economy. Based on Table 7, the number of households in Sumenep Regency is less in cluster 1 with high economic criteria. Households in Pamekasan Regency have fewer numbers for the highest economic criteria.

F. *Kruskal Wallis Test*

After obtaining the optimum cluster, then the average difference between clusters is tested. Tests are conducted to find out whether the results of the clustering have significantly different clusters. In testing differences between clusters on variables is a numerical scale. Tests were carried out in a nonparametric manner with the Kruskal Wallis test. This is done because the research variables used are not normally distributed per variable with the *Kolmogorov Smirnov* examination, and the results of the *mshapiro* test are the p-value are $2{,}2\times10^{-16}$ so that the data is not normally distributed. The *Kruskal Wallis* test was used on three numerical scale research variables. The hypothesis used in the *Kruskal Wallis* test is as follows.

H_0: there is no difference in average between clusters
H_1: there is a difference in average between clusters

The Kruskal Wallis test results on the research variables are represented in Table 8.

Table 8 The Result Kruskal Wallis Test to 5 Clusters

Variable	*Chi-Square*	df	*Chi-Square* Table
X_1	1408,271	4	9,49
X_2	60,302	4	9,49
X_3	442,726	4	9,49

The Kruskal Wallis test uses α value of 0.05. From the tests in Table 8, the three observation variables used include the chi-square value which is greater than the chi-square table, so it can be concluded that the three variables are significantly different between households in the five clusters.

V. Conclusion

After analysis and discussion on clustering of household in Madura using three methods, it can be concluded that the results of numerical and categorical data characteristics obtained an average total expenditure of 2,414 million and most of the sources of water used were sources of water that could not be purchased with the part of the water supply system uses piped system/public hydrants. Most households on the island of Madura do not pay for water from PAM (local water company).

Clustering numerical data uses the *agglomerative ward* hierarchical methods with two formed based on the results of the *pseudo-f value*, with each criterion being a high economy (group 1) and a low economy (group 2). With the number of households living in the economic criteria as many as 1,782 households and low economic criteria as many as 1,315 households. The group results with the highest number of households with low economic criteria are households that live in Sumenep Regency.

In the results of categorical scale data clustering using the k-modes method, three clusters are formed with each cluster criteria being low, medium and high. The number of households in cluster 1, cluster 2 and cluster 3 were 1975, 723 and 399 households respectively. The results of the group with the highest number of households with low economic characteristics are households that live in Pamekasan Regency.

The clustering of mixed-scale data namely numerical and categorical using the SWFM method formed five groups based on the value of the *Sw* and *Sb* ratio, with the criteria being a high economy, medium economy, the lowest economy, the highest economy, and a low economy. The results of the cluster with the highest number of households with low and very low economic characteristics were households living in Pamekasan and Sumenep Regencies.

REFERENCES

[1] M. P. Todaro, Pengembangan Ekonomi, Jakarta: Bumi Aksara, 2000.

[2] ESCAP-UNDP, ESCAP-UNDP Initiative for The Achievement of Millenium Development Goals in Asia And The Pacific., Bangkok: UN ESCAP, 2000.

[3] S. Suparman, "Kaum Perempuan Paling Peduli," *Media Informasi Air Minum dan Penyehatan Lingkungan*, pp. 9-11, 2007.

[4] BPPSPAM, Buku Kinerja PDAM, Jakarta: Kementrian Pekerjaan Umum dan Perumahan Rakyat, 2017.

[5] R. Walpole, Intoduction to Statistics, New York: Macmillan Publishing Co. Inc, 1995.

[6] B. Simamora, Analisis Multivariat Pemasaran., Jakarta: Gramendia Pustaka Umum, 2005.

[7] R. Johnson and D. Wichern, Applied Multivariate Statistical Analysis, New Jersey: Pearson Education, 2007.

[8] S. Sharma, Applied Multivariate Techniques, Canada: John Wiley & Sons, Inc., 1996.

[9] Halkidi, Batistakis and Vizirgiannis, "On Clustering Validation Techniques," *Journal of Intelligent Systems,* vol. 17, no. 2, pp. 107-145, 2001.

[10] L. Kaufman and P. J. Rousseeuw, Finding Groups in Data-An Introduction to Cluster Analysis, New York: Wiley, 1990.

[11] M. Reddy and B. Kavitha, "Clustering the Mixed Numerical and Categorical Dataset using Similarity Weight and Filter Method," *International Journal of Database Theory and Application,* vol. 5, no. 1, 2012 .

[12] M. Bunkers and R. M. James, "Definition of Climate Regions in the Northern Plains Using an Objective Cluster Modification Technique," *Journal Climate,* 1996.

[13] S. Winarna, Analisis Konsumsi Air Bersih Pelanggan Rumahtangga Berdasarkan Faktor-faktor yang Mempengaruhinya, Semarang: Universitas Diponegoro, 2003.

[14] W. W. Daniel, Statistik Nonparametrik Terapan, Jakarta: PT. Gramedia, 1989.

[15] S. Joachim and T. Hillenbrand, Determinants of Residential Water Demand in Germany. Working Paper Sustainability and Innovation No. S 3, 2007.

[16] P. S. Darr and C. Kamen, The Demand for Urban Water. Martinus Nijhoff Social Division, Leiden, 1976.