A Study Comparative of K-mean, K-medoids, and Fuzzy C-mean for Clustering Laboratories in ITS

Risky Frasetio Wahyu Pratama, Bahagiati Maghfiroh, dan Febrian Kristianda
Statistics Department, Faculty Matemathics Computation and Data Science,
Sepuluh Nopember Institute of Technology (ITS)
Jl. Arief Rahman Hakim, Surabaya 60111 Indonesia
email : riskyfrasetio92@gmail.com, bahagia.ti@gmail.com, febrian.kristianda@gmail.com

***Abstract*— Clustering is used extensively in various fields to obtain information from data (anomalies detection and identifying important features). One of the most commonly used clustering procedures is partitional-based clustering. In the is partitional-based clustering, number of clustering or partition is predefined. In this study, K-mean, K-medoid and Fuzzy C-Mean (FCM) were compared in clustering ITS laboratory . Simulation study is also used in order to see the clustering result in some settings of data scenarios. internal disspersion rate and pseudoF are used as index validity. The result of simulation study is k-mean generally give more optimal result than others for clustering data without outlier. Contrasly, if the outlier asserted, k-medoid perform better. The result of clustering ITS laboratory gives optimal number cluster k=5.**

.

Key Word—Cluster,K-Mean, K-Medoid, Fuzzy C Means, Fuzzy C-Medoid, Polynomial Fuzzy C-Means.

I. INTRODUCTION

Clustering is the procces of classifying objects into different groups based on information obtained from data that explains the relationship between objects with principles to maximize the similarity between members of one class and minimize the similarity between classes. Primary goals of clustering include getting information from data (anomalies detection and identifying important features), and classifying data[3].

Clustering is broadly divided into hierarchical and non-hierarchical methods [8]. Hierarchical algorithms recursively find nested clusters either in a top-down (divisive) or bottom-up (agglomerative) fashion. This method is often computationally ineficient[1]. Non-hierarchical method is divided into several procedures. One of the easiest is partitinional methods (ie k-mean and k-medoid). Partitional algorithms find all the clusters simultaneously as a partition of the data and do not impose a hierarchical structure[3].

k-means clustering method is the probably the most well known as a partitional procedure. The algorithm starts with initial centers, one for each cluster. This center is called centroid. All the instances are then compared with this centroid by a distance (euclidean, manhattan, etc) and assigned to the closest centroid. In the next stage, new centroids for each cluster are computed by using average of vector of the objects assigned to the cluster. This procedure is repeated until the members of each cluster are unchanged. Because of using mean vector as the centroid, this method is extremely sensitive to outliers[4]. The acurracy of k-means procedure is also very dependent upon the choice of initial centroids so that this method is sensitive with choosing the initial centroids[10]. K-medoid clustering is sometimes used, where representative objects called medoids are considered instead of centroids in order to get more robust result. Among many algorithm for k-medoids clustering, Partitioning Around Medoids (PAM) is known to be most powerful. Another approach that deals with outliers is C-means (FCM) algorithm by[2] that extend hard C-means clustering methods[6]. The aim of this study is to compare k-mean, k-medoid, and FCM performance by using simulated data with saveral scenario and implementation on a real data. In this study, we use k-mean, k-medoid, and FCM for grouping laboratories at Institut Teknologi of Sepuluh Nopember to evaluate the productivity of research development within each laboratory.

II. BACKGROUND THEORY

A. K-Mean Cluster

The K-means clustering algorithm is described in detail by [3]. An efficient version of the algorithm is presented here. The aim of the K-means algorithm is to divide M points in N dimensions into K clusters so that he within-cluster sum of squares is minimized.I t is not practical to require that the solution has minimal sum of squares against all partitions, except when M , N are small and K = 2. We seek instead" local" optima, solutions such that no movement of a point from one cluster to another will reduce the within-cluster sum of squares. Here is the pseudocode from[7] describing the iterations[12]:

1. Choose the number of clusters
2. Choose the metric to use
3. Choose the method to pick initial centroids
4. Assign initial centroid
5. Assign cases to closest centroid
6. Calculate centroids
7. For j ≤ number cluster, if centroid j was updated in the last iteration
    a. Calculate sum of square within cluster
    b. For i ≤ number cases in cluster
        i. Compute sum of square within cluster k ≠ j if case included.
        ii. If sum of square within cluster k < sum of square within cluster j, case change cluster

B. K-Medoids Cluster

In the k-medoids algorithm, rather than calculating the mean of the items in each cluster, a representative item, or medoid, is chosen for each cluster at each iteration. Medoids for each cluster are calculated by finding object i within the cluster that minimizes sum of square within cluster [4].

C. Fuzzy C-Mean

In fuzzy *k*-means clustering[2], each case has a set of degree of belonging relative to all clusters. It differs from previously presented *k*-means clustering where each case belongs only to one cluster at a time. In this algorithm, the centroid of a cluster (c*k*) is the mean of all cases in the dataset, weighted by their degree of belonging to the cluster (ω_k).

$$c_k = \frac{\sum_i \omega_k(x_i)x_i}{\sum_i \omega_k(x_i).}$$

The degree of belonging is a function of the distance of the case from the centroid, which includes a parameter controlling for the highest weight given to the closest case. It iterates until a user-set criterion is reached. Like the *k*-means clustering technique, this technique is also sensitive to initial clusters and local minima. It is particularly useful for dataset coming from area of research where partial belonging to classes is supported by theory.

D. Validity Clustering

The performance of the algorithm was evaluated by the average percentage of correct classification (recovery rate) and the internal cluster dispersion rate of the final partition defined as[11]

$$ICD = 1 - \frac{SSB}{SST} = 1 - R^2$$

$SSB = \sum_{j=1}^{k} d_{j0}^2$ $SST = \sum_{\ell=1}^{n} d_{\ell}^2$. d_{j0}^2 is the Euclidean distance between the jth cluster center vector and the overall sample mean vector, d_{ℓ}^2 is the Euclidean distance between the l th observation vector and the overall sample mean vector, k is the number of clusters, n is the number of observed vectors.

Another validity method used to determine the number of optimum groups is Pseudo F-statistic. The highest Pseudo F indicates that the group shows optimal results, where the diversity in the group is very homogeneous whereas the intergroup is very heterogeneous. Here's the formula of Pseudo F [9]

$$PseudoF = \frac{R^2/(c-1)}{(1-R^2)/(n-c)}$$

III. METHODOLOGY

For comparative study, simulation study with some scenario settings was used in this study. The second stage is the application of data on laboratory achievement in ITS. Scenario setting in simulation study uses combination of observation number (N): 50 observation without outliers, 500 observation without outlier and 200 observation with adding outliers (outlier ratio is 0,3 from 200); number of clusters (k): [2,3,4,5]; number of variables (p): [5,10,20,40]; the process of generating data is implemented by using R clusterGeneration package by Qu and Joe (2015) that has implemented algorithm proposed by [10] for generating cluster population. Each scenario of simulation setting was analyzed using K-Mean, K-Medoid, and Fuzzy K-Mean methods to see the best performance by using icd rate and pseudoF validity index.

Furthermore, there is a grouping analysis on laboratory capability data available in ITS using the three clustering methods. In order to obtain the best clusters, it can be concluded which laboratory group has a high index of research achievement. Stages of analysis is performed starting from data preprocessing (i.e scalling data into range 0 1 and using imputation to treat missing value using K-Nearest Neighbour) . 5 nearest neighbour is defined by using laboratory neighbours in the same faculty rather than use all faculty. Furthermore, factor analysis was done to reduce the dimension of the variable and to group variables with high correlation in order to get easier interpretation of the clustering result. The next step is grouping the laboratory to assess the research achievements of each laboratory and then interpret the result. All the analysis is conducted by using R 3.3.0.

IV. RESULT

A. Simulation Study

Simulation study is used to see clustering performance among all methods on the saveral scenario simulation.

Table 1. Simulation of Observation Without Outlier

Number of Variabel	Number of Observation	Clustering Mrthod	Number of Cluster			
			2		3	
			ICD	Pseudo F	ICD	PseudoF
5	50 (Without Outlier)	K-Mean	0.5522	38.928	0.4088	33.985
		K-Medoid	0.5281	42.885	0.4602	27.57
		Fuzzy K-Mean	0.5524	38.895	0.409	32.205
	500 (Without Outlier)	K-Mean	0.5817	358.18	0.46076	290.83
		K-Medoid	0.6009	330.75	0.4591	292.82
		Fuzzy K-Mean	0.5817	358.08	0.4608	290.8
10	50 (Without Outlier)	K-Mean	0.6394	27.071	0.6573	12.254
		K-Medoid	0.5785	34.975	0.64955	12.679
		Fuzzy K-Mean	0.6402	26.982	0.6638	11.901
	500 (Without Outlier)	K-Mean	0.7203	193.4	0.6579	129.23
		K-Medoid	0.7145	198.98	0.6564	130.06
		Fuzzy K-Mean	0.7203	191.11	0.658	129.16
20	50 (Without Outlier)	K-Mean	0.8224	10.367	0.7299	8.6952
		K-Medoid	0.8372	9.3318	0.6936	10.382
		Fuzzy K-Mean	0.8243	10.231	0.7343	8.5041
	500 (Without Outlier)	K-Mean	0.8353	98.174	0.77757	71.087
		K-Medoid	0.8458	90.825	0.803	60.979
		Fuzzy K-Mean	0.8353	98.131	0.7779	70.957
40	50 (Without Outlier)	K-Mean	0.8902	5.9178	0.8408	4.4496
		K-Medoid	0.8539	8.2064	0.8342	4.6711
		Fuzzy K-Mean	0.8962	5.5613	0.8551	3.9829
	500 (Without Outlier)	K-Mean	0.9182	44.364	0.8719	36.507
		K-Medoid	0.8767	70.011	0.8619	39.811
		Fuzzy K-Mean	0.9189	43.931	0.8728	36.203

Table 1. Simulation of Observation Without Outlier (Continued)

Number of Variable	Number of Observation	Clustering Mrthod	Number of Cluster			
			4		5	
			ICD	PseudoF	ICD	Pseudo F
5	50 (Without Outlier)	K-Mean	0.449	18.814	0.4781	12.283
		K-Medoid	0.447	18.972	0.464	12.998
		Fuzzy K-Mean	0.4501	18.731	0.457	13.364
	500 (Without Outlier)	K-Mean	0.4068	241.07	0.4021	183.99
		K-Medoid	0.3943	253.94	0.4040	182.54
		Fuzzy K-Mean	0.4069	240.97	0.4022	183.9
10	50 (Without Outlier)	K-Mean	0.5855	10.855	0.5517	9.1401
		K-Medoid	0.6075	9.90873	0.5827	8.0561
		Fuzzy K-Mean	0.5886	10.716	0.5621	8.7628
	500 (Without Outlier)	K-Mean	0.577	121.21	0.5686	93.89
		K-Medoid	0.6233	99.921	0.5956	84.012
		Fuzzy K-Mean	0.5771	121.16	0.5688	93.795
20	50 (Without Outlier)	K-Mean	0.6902	6.8802	0.7201	4.3728
		K-Medoid	0.6977	6.6444	0.7310	4.1392
		Fuzzy K-Mean	0.6954	6.71492	0.706	4.6844
	500 (Without Outlier)	K-Mean	0.7597	52.311	0.7541	40.362
		K-Medoid	0.8071	39.493	0.7585	39.407
		Fuzzy K-Mean	0.7601	52.175	0.7549	40.169
40	50 (Without Outlier)	K-Mean	0.6902	6.8802	0.7722	3.3192
		K-Medoid	0.6977	6.6444	0.778	3.2104
		Fuzzy K-Mean	0.6954	6.7149	0.7741	3.2826

500 (Without Outlier)	K-Mean	0.8555	27.936	0.8319	25.003
	K-Medoid	0.8135	37.902	0.8355	24.365
	Fuzzy K-Mean	0.8572	27.551	0.8336	24.704

According to above table, in the case of without outliers asserted in data, the overall mean icd rate and mean pseudoF validity index obtained for k-medoid are generally better than others. But as number of observation increase, K-Mean method give the best result overally. Meanwhile if we look based on the number of variable, when the number of small variables (5) the best clustering method is K-mean, but when its variable increases (10, 20, 40), the clustering method that gives the best performance is K-Medoid. K-mean method has decreased performance as the number of variable is increased. The next simulation scenario, outliers are asserted.

Table 2. Simulation of With Observation Outlier

Number of Variable	Number of Observation	Clustering Mrthod	Number of Cluster			
			2		3	
			ICD	PseudoF	ICD	PseudoF
5	200 (Ratio Outlier 30%)	K-Mean	0.8969	29.645	0.7098	52.528
		K-Medoid	0.6925	114.58	0.6386	72.722
		Fuzzy K-Mean	0.7135	103.62	0.654	67.979
10	200 (Ratio Outlier 30%	K-Mean	0.841	48.772	0.7525	42.271
		K-Medoid	0.8494	45.732	0.7263	48.428
		Fuzzy K-Mean	0.842	48.398	0.7556	41.573
20	200 (Ratio Outlier 30%	K-Mean	0.897	29.626	0.8792	17.65
		K-Medoid	0.8683	39.121	0.8673	19.663
		Fuzzy K-Mean	0.9012	28.3	0.8928	15.433
40	200 (Ratio Outlier 30%	K-Mean	0.96	10.744	0.9395	8.2693
		K-Medoid	0.9454	14.899	0.9002	14.248
		Fuzzy K-Mean	0.9658	9.1355	0.949	6.9064

Table 2. Simulation of With Observation Outlier (Continued)

Number of Variable	Number of Observation	Clustering Mrthod	Number of Cluster			
			4		5	
			ICD	PseudoF	ICD	PseudoF
5	200 (Ratio Outlier 30%)	K-Mean	0.6169	52.992	0.5668	48.726
		K-Medoid	0.5911	59.04	0.5615	49.786
		Fuzzy K-Mean	0.6191	52.505	0.5681	48.472
10	200 (Ratio Outlier 30%	K-Mean	0.6954	37.375	0.7233	24.392
		K-Medoid	0.7078	35.22	0.6551	33.558
		Fuzzy K-Mean	0.7299	31.571	0.7182	21.781
20	200 (Ratio Outlier 30%	K-Mean	0.8052	20.643	0.8062	15.322
		K-Medoid	0.8244	18.173	0.7847	17.493
		Fuzzy K-Mean	0.8071	20.39	0.7994	16
40	200 (Ratio Outlier 30%	K-Mean	0.8942	10.094	0.8506	11.199
		K-Medoid	0.9147	7.9595	0.8224	13.769
		Fuzzy K-Mean	0.904	9.0617	0.8603	10.356

Based on table 2, As a whole K-Medoid look better than others. This is because k-medoid is a robust method when the observation contains outliers.

B. Laboratory Data

Prior to analysis, preprocessing data is performed to reduce noise in the data, such as scalling, missing and outlier data. This needs to be addressed first so that the results are smoother and represent the conditions in the field well.

Scaling is done because the inter-variables have different denomination. So it should be standardized by scalling, which has a range of 0-1.

Missing data is overcome by using the K-Nearest Neighbor method with K = 5 and the neighbor in the faculty itself. Per Jönsson and Claes Wohlin (2004) suggest that by relaxing the method rule with respect to neighboring selection, the method's capability remains high for large amounts of lost data without affecting the quality of imputation. In some instances, the nearest neighboring imputation method provides unbiased and consistent asymptotic estimators of population (or total) mean function, population distribution, and population number[4]. In this study there are three laboratories in the Faculty of Industrial Technology which are not included in the analysis due to unavailability of data, they are the Laboratory of Mineral Processing Technology and Materials of Materials Engineering and Metallurgy, Development of Industrial Systems and Management majoring in Industrial Engineering and Chemical Reaction Engineering Department of Chemical Engineering.

Table 3. Number of Data Missing Each Faculty

FTSP	FTIF	FBMT	FV	FTI	FTE
103	6	0	0	25	0

Outlier data has been accommodated by the medoids method based on the median data. The median is a robust measure of outliers compared to the mean.

Furthermore, factor analysis was done to reduce the variable dimension quite a lot. Hoping it can be interpreted easily. Here's the result of factor analysis,

Table 4. Grouping Variables with Factor Analysis

Faktor	No	Information
Activities Laboratory	11	Community service activities that are Chosen by the relevant Laboratory Lecturer
	14	Lecturers' involvement in international consortium / research forum
	15	Lecturers' involvement in consortium / national research forum
	18	Modules / textbooks developed in the current period
	19	Studies with Partners from PT Overseas
	21	Publications in National Journal
	24	Lecturers Writing Textbook BerISBN and Distributed in Market (accumulation)
	25	Lecturers Invited as Invited Speakers at International Seminar
	26	Lecturers who attended the training / workshop
	28	Number of Lecturers who become Editor or Reviewer International Journal
	29	IPRs registered
	32	LBE certificates
Citation	8	Index H in Scopus
	22	Lecturer Criteria in Google Scholar
	23	Citation Lecturers in Scopus
	27	Lecturers who become Members of the Association of International Professions
Profile	3	Lecturers of Laboratory Members (including Chairman)
	5	Research grant (Rp Million)
	6	Research grant funded by related Laboratory Lecturer (Rp Million)
	9	Research Titles
	10	Research titles Headed by Lecturers from relevant Laboratory
	12	Classes / lab courses run / serviced by the Laboratory
	16	Students involved in Lecturer's research
International Publication	17	Students involved in dedication of Lecturer
	20	Publications in International Journal Indexed Scopus, Thomson etc.
	30	International Co-Authorship
	31	Publications at International Seminar
Cooperation Research	4	Fund dedication to the community that is Chosen by the related Laboratory Lecturer (Rp Million)
	7	Fund of Research Cooperation and PPM with Institution / Industry headed by Laboratory Lecturer related (Rp Million)
	13	Research Cooperation and PPM with Institution / Industry per year which is chaired by related Laboratory Lecturer
	33	International journal publications

Exploration of data was conducted to find out the characteristics of each Exploration of data was conducted to find out the characteristics of each variable and comparison between faculties. Showed from the figure for all factors, FTI has the highest value almost in all variables. Because FTI has the most majors and the largest number of labs. Unlike the Faculty of FBMT classified as a new faculty with only 1 department and 2 labs. Neither is for other faculty Graph of comparison of the characteristics of each variable and comparison between the faculty can be seen in the appendix.

Table 5 shows the summaries of icd rate and pseudoF comparison for every possible cluster (2:10) within each clustering method

Table 5. Comparison of Validity Clusters from Laboratory Performance Data

Clustering Method	Number of Cluster					
	1		2		3	
	ICD	PseudoF	ICD	PseudoF	ICD	PseudoF
K-Mean	0.87398	14.85114	0.71456	20.37258	0.598039	22.62848
K-Medoid	0.833849	20.5236	0.488616	12.59751	0.666561	16.84137
Fuzzy K-Mean (m = 1,1)	0.85346	17.6856	0.801918	20.32904	0.574154	24.97029
Fuzzy K-Mean (m = 1,5)	0.86603	15.93296	0.73122	18.74642	0.60152	22.30268
Fuzzy K-Mean (m = 1,9)	0.87125	15.2211	0.776422	14.68594	0.719885	13.10007

Table 5. Comparison of Validity Clusters from Laboratory Performance Data (Continued)

Metode Clustering	Number of Cluster					
	5		6		7	
	ICD	PseudoF	ICD	PseudoF	ICD	PseudoF
K-Mean	0.463468	28.94115	0.36997	33.71779	0.405323	23.9637
K-Medoid	0.555888	19.97312	0.488616	20.72258	0.381944	26.43041
Fuzzy K-Mean (m = 1,1)	0.532332	21.96314	0.428326	26.42649	0.405865	23.90991
Fuzzy K-Mean (m = 1,5)	0.589395	17.41638	0.451418	24.06178	0.43252	21.42989
Fuzzy K-Mean (m = 1,9)	0.577562	18.28539	0.56553	15.21143	0.474961	18.05549

Table 5. Comparison of Validity Clusters from Laboratory Performance Data (Continued)

Clustering Method	Number of Cluster					
	8		9		10	
	ICD	PseudoF	ICD	PseudoF	ICD	PseudoF
K-Mean	0.309608	30.89998	0.287076	29.80085	0.258495	30.27918
K-Medoid	0.350157	25.717	0.329338	24.43669	0.301886	24.4098
Fuzzy K-Mean (m = 1,1)	0.31139	30.64385	0.273236	31.91802	0.251319	31.44504
Fuzzy K-Mean (m = 1,5)	0.413662	19.64154	0.367449	20.65761	0.369655	17.99961
Fuzzy K-Mean (m = 1,9)	0.457189	16.45229	0.446149	14.89683	0.4333	13.80285

Optimal value of icd rate and pseudoF validity index respectively is 0,25132 for FCM (m=1,1) with 10 cluster and 33,7178 for k-means with 6 cluster. Generally, optimal modelling result are obtained by minimum average of icd rate and maximum average of pseudoF. As shown in table 5, k-mean give the optimal results with average of icd rate (0,47561) and average of pseudoF (26,1617).

Table 6. Mean ICD and PseudoF from the Number of Cluster

Clustering Method	ICD	PseudoF
K-Mean	**0,475614**	**26,16165**
K-Medoid	0,523351	21,29468
Fuzzy K-Mean dengan Parameter Fuzzy (m = 1,1)	0,482786	25,47682
Fuzzy K-Mean dengan Parameter Fuzzy (m = 1,5)	0,535875	19,79876
Fuzzy K-Mean dengan Parameter Fuzzy (m = 1,9)	0,591366	15,52349

By looking scree plot on figure1, the icd rate index of k-means clustering decreases sharply untill k=5, and then slowly decreases later on so that optimal number of cluster for grouping the laboratory is defined by k=5.

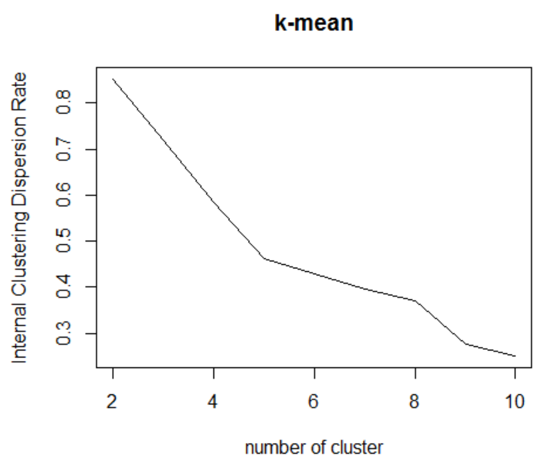

Figure 1. Plotting k optimal

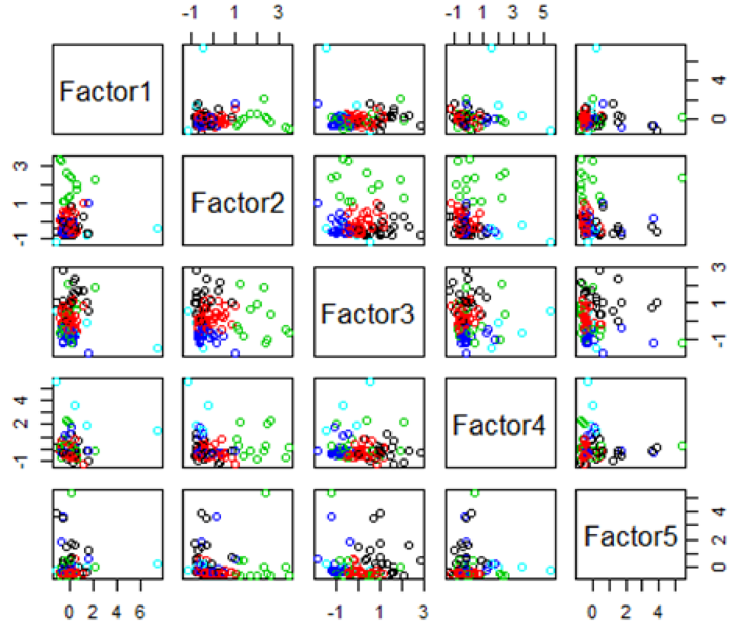

Figure 2. Variable in 3[rd] Factor

To simplify the interpretation of clustering result, Figure 2 was obtained to show the pairs plot of all factor conducted by 31 variables. As we can see that observations with high factor 2 values tend to be grouped into cluster 3 (green). Cluster 1 (black) tend to contain high values of factor 3 (Profile) with low values of factor 1 (Activities Laboratory), factor 2 (Citation), factor 4 (International Publication) and medium to high values of factor 5(Cooperation Research). Cluster 5 (light blue) tend to be containned by labratories with high values of factor 5.

V. CONCLUSION

The conclusions can be obtained from this research, the first when using the simulation data the case of without outliers asserted in data, the overall mean icd rate and mean pseudoF validity index obtained for k-medoid are generally better than others. But as number of observation increase, K-Mean method give the best result overally. Meanwhile if we look based on the number of variable, the clustering method that gives the best performance is K-Medoid. K-mean method has decreased performance as the number of variable is increased. The next simulation scenario, outliers are asserted. Mean while if any outliers in dataset As a whole K-Medoid look better than others. This is because k-medoid is a robust method when the observation contains outliers.

The second when applied to lab data Generally, optimal modelling result are obtained by minimum average of icd rate and maximum average of pseudoF. As shown in table 6, k-mean give the optimal results with average of icd rate (0,47561) and average of pseudoF (26,1617).

VI. APPENDIX

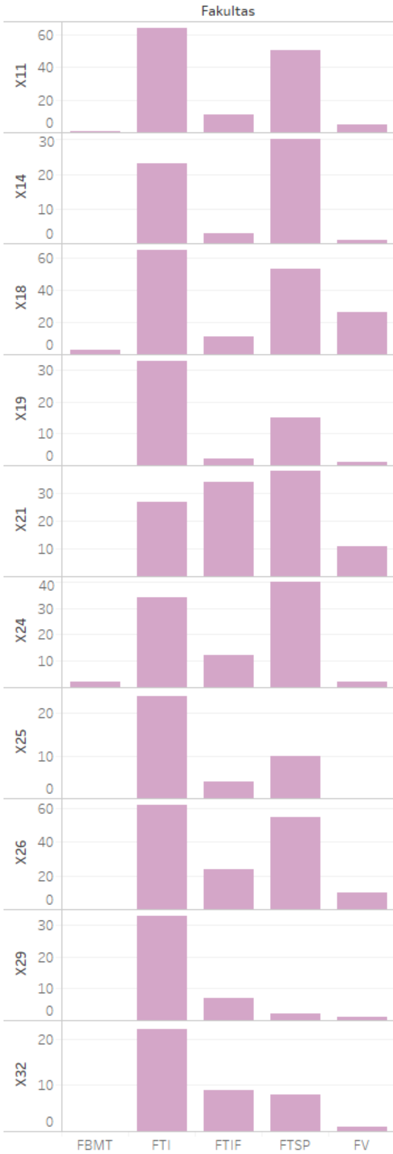

Figure 3. Variable in 1nd Factor

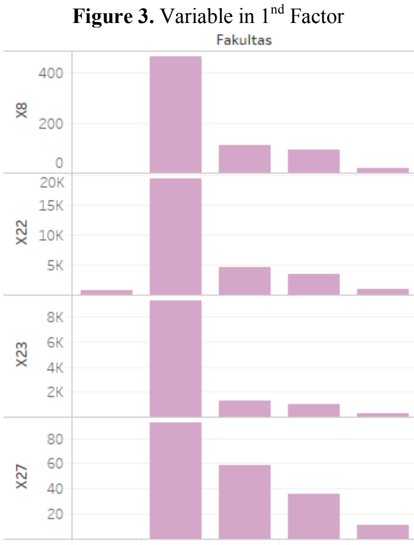

Figure 4. Variable in 2nd Factor

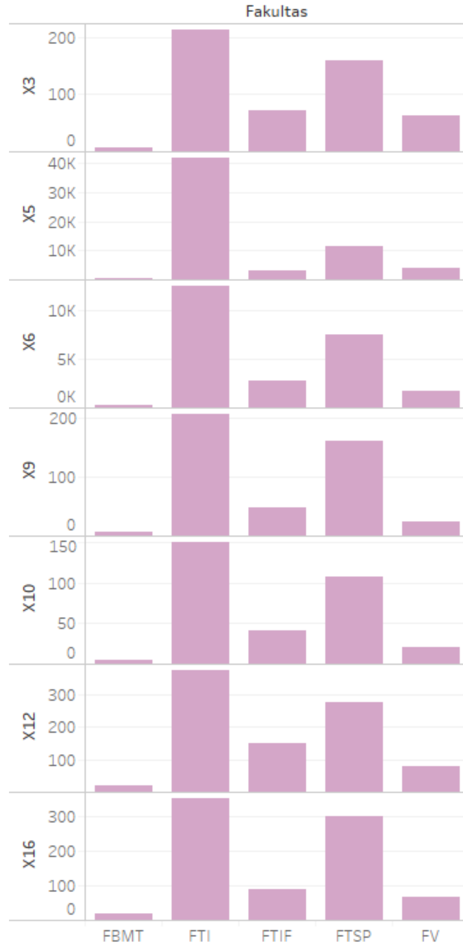

Figure 5. Variable in 3rd Factor

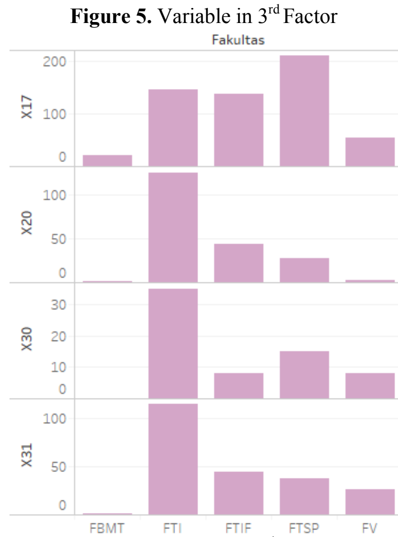

Figure 6. Variable in 4th Factor

Figure 7. Variable in 5th Factor

BIBLIOGRAPHY

[1] Bataineh, K. M., Naji, M., & Saqer, M. (2011). A Comparison Study between Various Fuzzy Clustering Algorithms. *Jordan Journal of Mechanical & Industrial Engineering*, *5*(4).

[2] Bezdek, J. C., Coray, C., Gunderson, R., & Watson, J. (1981). Detection and characterization of cluster substructure i. linear structure: Fuzzy c-lines. *SIAM Journal on Applied Mathematics*, *40*(2), 339-357.

[3] Celebi, M. E., Kingravi, H. A., & Vela, P. A. (2013). A comparative study of efficient initialization methods for the k-means clustering algorithm. *Expert systems with applications*, *40*(1), 200-210.

[4] Chawla, S., & Gionis, A. (2013, May). k-means–: A unified approach to clustering and outlier detection. In *Proceedings of the 2013 SIAM International Conference on Data Mining* (pp. 189-197). Society for

[5] Chen, J., & Shao, J. (2000). Nearest neighbor imputation for survey data. *Journal of official statistics*, *16*(2), 113. Industrial and Applied Mathematics.

[6] Ghosh, S., & Dubey, S. K. (2013). Comparative analysis of k-means and fuzzy c-means algorithms. *International Journal of Advanced Computer Science and Applications*, *4*(4), 35-39.

[7] Hartigan, J. A. and Wong, M. A. (1979). A K-means clustering algorithm. *Applied Statistics* **28**, 100–108.

[8] Johnson R.A dan Wichern D.W, 2007. *Applied Multivariate Statistical Analysis,6thed*. Prentice Hall International Inc, New Jersey.

[9] Kartidan, H. S., & Irhamah, I. (2013). Pengelompokan Kabupaten/Kota di Provinsi Jawa Timur Berdasarkan Indikator Pendidikan SMA/SMK/MA dengan Metode C-Means dan Fuzzy C-Means. *Jurnal Sains dan Seni ITS*, *2*(2), D288-D293.

[10] Milligan, G.W., Cooper, M.C., 1980. An examination of the effect of six types of error perturbation on fifteen clustering algorithms. Psychometrika 45 (3), 159–179.

[11] Mingoti, S. A., & Lima, J. O. (2006). Comparing SOM neural network with Fuzzy c-means, K-means and traditional hierarchical clustering algorithms. *European Journal of Operational Research*, *174*(3), 1742-1759.

[12] Morissette, L., & Chartier, S. (2013). The k-means clustering technique: General considerations and implementation in Mathematica. *Tutorials in Quantitative Methods for Psychology*, *9*(1), 15-24.