Indonesian Journal of Applied Statistics ISSN 2621–086 X
Volume 3 No. 1 May 2020

Laboratory Clustering using *K*-Means, *K*-Medoids, and Model-Based Clustering

Niswatul Qona'ah[1], Alvita Rachma Devi[2], and I Made Gde Meranggi Dana[3]
[1]Statistics Study Program, Universitas Sebelas Maret
[2,3]Statistic Department, Institut Teknologi Sepuluh Nopember

[1]niswatulqonaah@staff.uns.ac.id

Abstract. Institut Teknologi Sepuluh Nopember (ITS) is an institute which has about 100 laboratories to support some academic activity like teaching, research and society service. This study is clustering the laboratory in ITS based on the productivity of laboratory in carrying out its function. The methods used to group laboratory are *K*-Means, *K*-Medoids, and Model-Based Clustering. *K*-Means and *K*-Medoids are non-hierarchy clustering method that the number of cluster can be given at first. The difference of them is datapoints that selected by *K*-Medoids as centers (medoids or exemplars) and mean for *K*-Means. Whereas, Model-Based Clustering is a clustering method that takes into account statistical models. This statistical method is quite developed and considered to have advantages over other classical method. Comparison of each cluster method using Integrated Convergent Divergent Random (ICDR). The best method based on ICDR is Model-Based Clustering.

Keywords : *K*-Means, *K*-Medoids, Laboratory, Model-Based Clustering

1. Introduction

In the education and teaching at university or institute, laboratory is used for teaching activities that require specific skill practice or direct experiences for learners. The laboratory also serves to conduct research, both using experimental methods and laboratory surveys. The laboratory can also be used to support society service programs. According to article 3 paragraph 4 of PP 60/1999, "Society service is an activity that utilizes science in order to contribute to the progress of society" [1].

Institut Teknologi Sepuluh Nopember (ITS) has about 100 laboratories to support teaching, research, and society service activites. This research is clustering the laboratory in ITS based on productivity and activeness in carrying out its function. This clustering aims to find out laboratory group which have high productivity. This group can be used as role models for other laboratories. In fact, it can also be known which laboratories still have low productivity, so it can be prioritized to be improved both resources and facilities.

The methods used for laboratories clustering are *K*-Means, *K*-Medoids and Model-Based Clustering. *K*-Means is a classical clustering method that the number of cluster can be given at first. This method is clustering the observation which have similarity of characteristics into same cluster, and the observation which have difference

of characteristics in to another cluster. A robust alternative to *k*-means is *k*-medoids, which is based on medoids. Model-Based Clustering is a clustering method that honor statistical model. This clustering method is quite developed and considered to have advantages over other classical method [2]. This study will compare the result of laboratories clustering in ITS using *K*-Means, *K*-Medoids and Model-Based Clustering method.

2. Literature

2.1. Laboratory. Laboratory is an important element and one of the requirements for the existence of college. In article 56 paragraph 1 of PP No. 60/1999 stated, "Every university/institute must have library, computer center, laboratory/studio, and other supporting element required for the organization of university".

Laboratory is a place of Tri Dharma Higher Education: education and teaching, research, and society service. Availability of facilities including laboratories is one of the main aspects performance indicators in the preparation of institutional portofolio to be assessed by the National Accreditation Board of Higher Education (BAN-PT). To complete normative performance standards, the availibity of existing facilities should be directed to the achievement of the vision, mission, and the purpose of institution, also the feasibility and compatibility of facilities to support and has high access at education and the other Tri Darma activity [1].

2.2. Factor Analysis. Factor analysis is multivariate method to analyze the independence of variable which has high dimension, with goal of factorization. The purpose of factor analysis is to illustrate the covariance relationships between some underlying but unobserved variables, random quantities called factors [3]. A random vector is observed X with a p component, having an average of μ and a covariance matrix. Factor analysis model is as follows:

$$\begin{aligned} x_1 - \mu_1 &= l_{11}F_1 + l_{12}F_2 + ... + l_{1m}F_m + \varepsilon_1 \\ &\vdots \\ x_p - \mu_p &= l_{p1}F_1 + l_{p2}F_2 + ... + l_{pm}F_m + \varepsilon_p \end{aligned} \quad (1)$$

where

μ : mean of each variable

ε : specific factor

F : common factor

l : loading factor

The purpose of factor analysis is:

1. Reduced (with no loss of information) 1 set of origin variables into 1 set of new variables (Principal Component) of smaller dimension, based on interdependence structure (correlation /covariance between variables). In such a way that each PC can explain as much data variability as possible and between independent PC.
2. It is easier to interpret the structure of the covariance relationship between the origin variables, in 1 set of new variables.

2.3. Non-Hierarchy Clustering Method. Non-Hierarchy clustering method is used when the number of cluster has been known and this method usually used for clustering the large size data. The methods that belong to Non-Hierarchy clustering method are *k*-means and *k*-medoids.

2.3.1 *K*-Means. *K*-Means is one of the non-hierarchical data clustering methods that divide data into one or more cluster. Data that have the same characteristics are grouped into one same cluster, whereas those with different characteristics are grouped in another cluster.

The clustering process begins by identifying the data to be grouped $X_{ij}(i=1,...,n;\ j=1,...,m)$ with n is the amount of data to be grouped and m is the number of variables. At the beginning of the iteration, the center of each group is determined arbitrarily, $C_{kj}(k=1,...,K;\ j=1,...,m)$. Then calculated the distance between each data with each center group. To calculate the distance of *i*-th data (x_i) at the center of the *k*-cluster (c_k), named (d_{ik}), can be used Euclidian formula, as in equation (2).

$$d_{ik} = \sqrt{\sum_{j=1}^{m}(x_{ij} - c_{ij})^2} \tag{2}$$

A data will be a member of the *k*-th cluster if the distance of the data to the center of the *k*-cluster is of least when compared to the distance to the center of the other group. This can be calculated using equation (3).

$$\min \sum_{k=1}^{K} d_{ik} = \sqrt{\sum_{j=1}^{m}(x_{ij} - c_{ij})^2} \tag{3}$$

The new group center value can be calculated by finding the average value of the data belonging to the group, using equation (3).

$$c_{ij} = \frac{\sum_{i=1}^{p} x_{ij}}{p} \tag{4}$$

where $x_{ij} \in k$-cluster dan p is the number of k-cluster member [3].

2.3.2 *K*-Medoids. The *k*-medoids algorithm is a clustering algorithm related to the *k*-means algorithm and the medoid shift algorithm. Both the *k*-means and *k*-medoids algorithms are partitional and both attempt to minimize the distance between points labeled to be in a cluster and a point designated as the center of that cluster. In contrast to the *k*-means algorithm, *k*-medoids chooses datapoints as centers (medoids or exemplars) and works with a generalization of the Manhattan Norm to define distance between data points.

Suppose that n objects having p variables each should be grouped into k ($k < n$) clusters, where k is assumed to be given. Let j-th variable of object i as X_{ij} $(i = 1,...,n; j = 1,..., p)$. The Euclidean distance will be used as a dissimilarity measure in this study although other measures can be adopted. The Euclidean distance between object i and object j is given by

$$d_{ij} = \sqrt{\sum_{a=1}^{p} (X_{ia} - X_{ja})^2} \quad i = 1,...,n; j = 1,...,n \tag{5}$$

It should be noted that the above Euclidean distance will be adopted in *K*-means and PAM algorithms in this study. The proposed algorithm is composed of the following three steps [2].

Step 1: (Select initial medoids)

1-1. Calculate the distance between every pair of all objects based on the chosen dissimilarity measure (Euclidean distance in our case).

1-2. Calculate v_j for object j as follows:

$$v_j = \sum_{i=1}^{n} \frac{d_{ij}}{\sum_{l=1}^{n} d_{il}}, \; j = 1,...,n$$

1-3. Sort v_j in ascending order. Select k object having the first k smallest values as initial medoids.

1-4. Obtain the initial cluster result by assigning each object to the nearest medoid.

1-5. Calculate the sum of distances from all objects to their medoids.

Step 2: (Update medoids)

Find a new medoid of each cluster, which is the object minimizing the total distance to other objects in its cluster. Update the current medoid in each cluster by replacing with the new medoid.

Step 3: (Assign objects to medoids)

3-1. Assign each object to the nearest medoid and obtain the cluster result.

3-2. Calculate the sum of distance from all objects to their medoids. If the sum is equal to the previous one, then stop the algorithm. Otherwise, go back to the Step 2.

2.4. Model-Based Clustering. Model-Based Clustering (MBC) is a grouping method that takes into account statistical models. This model was first applied in [4] for grouping objects in the population. The assumption used in MBC is that in a population a subpopulation can be taken that has a certain probability distribution and each subpopulation has different parameters. All subpopulations have a distribution of mixture opportunities with different proportions for each subpopulation. This assumption leads to the mathematical probability model of the finite mixture model. Currently the use of finite mixture in clustering has grown rapidly and became one of the popular clustering methods.

The MBC framework was developed using the eigenvalue of the variance and covariance matrix $(\sum_g)$ as follows:

$$(\sum_g) = \lambda_g \mathrm{D}_g \mathrm{A}_g \mathrm{D}_g^{\mathrm{T}} \tag{6}$$

where :

λ_g is a scalar value indicating the ellipse volume.

D_g is an orthogonal eigenvector matrix which is the orientation of principal component $\sum_g$.

A_g is a diagonal matrix with proportional elements on eigenvalue and shows the contours of its density function.

3. Results and Discussion

There are 100 laboratories in ITS that will be grouped based on 31 variables which is an indicator of laboratory productivity. However, these 31 variables have similarities between the variables with each other, so it needs to be done variable reduction. In this analysis, variable reduction using factor analysis.

3.1. Factor Analysis. The 31 variables of laboratory productivity indicators can be reduced by factor analysis. Determination of the number of common factors can be explained through the scree plot as follows.

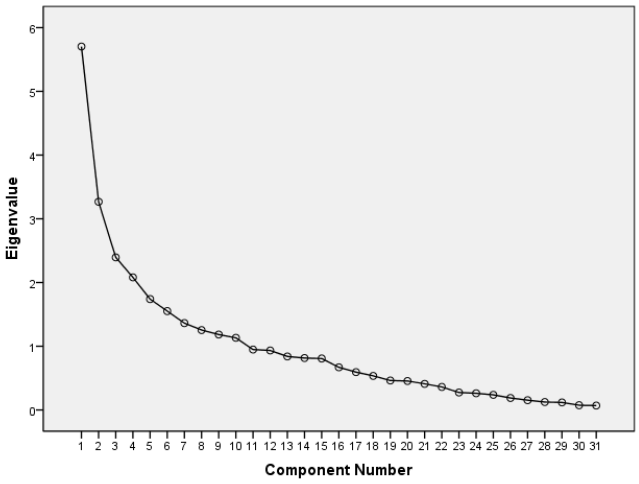


Figure 1. Scree Plot

In Figure 1 it can be observed that the steepest slope chart change occurs from point 1 to point 10. This shows that the number of common factors taken is 10 factors. To know how many variations that can be explained by the 10 factors formed then can be seen in Table 1.

Table 1. Total Variance Explained

Component	Initial Eigen Value			Component	Initial Eigen Value		
	Total	% of variance	Cumulative %		Total	% of variance	Cumulative %
1	3.394	10.950	10.950	6	1.750	5.647	50.165
2	3.349	10.802	21.751	7	1.721	5.550	55.716
3	2.407	7.763	29.514	8	1.669	5.383	61.099
4	2.330	7.516	37.030	9	1.472	4.749	65.848
5	2.321	7.489	44.519	10	1.259	4.060	69.908

Table 1 shows that the first component (factor 1) to the tenth component (factor 10) has more than one eigenvalue. Hence, 10 common factors can be calculated which can account for 69.908% of the overall variability of the data. Thus, in the next analysis

used 10 factors formed to classify the laboratory. The name of the factors and the variables that constitute them are presented in Table 2.

Table 2. The Name of The Factors Obtained

Factor	Variable	The Name of Factor	Factor	Variable	The Name of Factor
1	X4	research productivity	5	X2	Activity of lab lecturer in research
1	X7	research productivity	5	X5	Activity of lab lecturer in research
1	X8	research productivity	5	X11	Activity of lab lecturer in research
1	X9	research productivity	5	X22	Activity of lab lecturer in research
1	X14	research productivity	6	X1	Activity lab in teaching activities
2	X6	lecturer productivity in journal writing	6	X10	Activity lab in teaching activities
2	X20	lecturer productivity in journal writing	7	X13	The lecturer's activity in the study forum
2	X21	lecturer productivity in journal writing	7	X24	The lecturer's activity in the study forum
2	X26	lecturer productivity in journal writing	7	X28	The lecturer's activity in the study forum
3	X12	Activity Relationships with outsiders	8	X19	Productivity works
3	X17	Activity Relationships with outsiders	8	X30	Productivity works
3	X23	Activity Relationships with outsiders	9	X3	Activeness in science development
4	X18	Journal publication	9	X16	Activeness in science development
4	X29	Journal publication	10	X15	Activity of student profession and lecturer
4	X31	Journal publication	10	X25	Activity of student profession and lecturer

3.2. *K*-Means. *K*-means clustering is the most commonly used unsupervised machine learning algorithm for clustering data set into a set of *k* groups/cluster based on their similarity, where *k* represents the number of cluster. In k-means clustering, each cluster is represented by its center or centroid which corresponds to the mean of points assigned to the cluster. The first step is determine the number of clusters (*k*).

Because the purpose of this paper to cluster given data set into 2 groups, so we set k as 2. Using *k*-means algorithms with *k*=2, the result is 75 observations are allocated

 ISSN 2621–086 X

in cluster 1 and 25 observations in cluster 2. The visualization data in a scatter plot according to its cluster assignment can be seen in Figure 2.

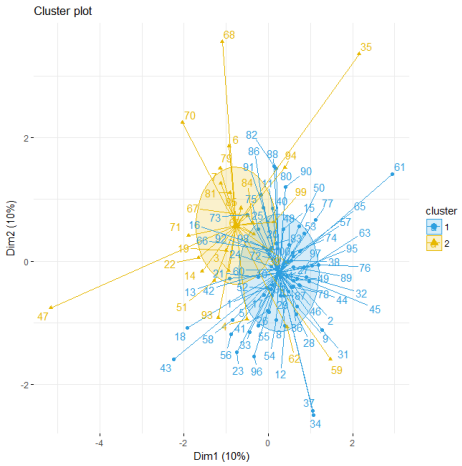
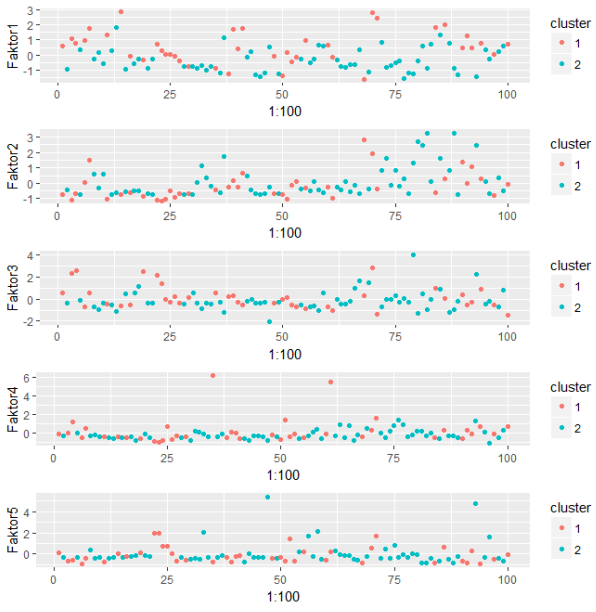


Figure 2. The Result of Laboratory Clustering using *K*-Means

The blue point represent lab which was clustered into cluster 1, meanwhile the yellow one is the number of lab which was put in cluster 2. Because the data contain 10 variables, visualize the data in a scatter plot with 2 dimension can be tricky. The plot above only explain 20% of point variability so it can't explain what each cluster mean to data. Furthermore, we can see cluster in each factor/variables in Figure 3.



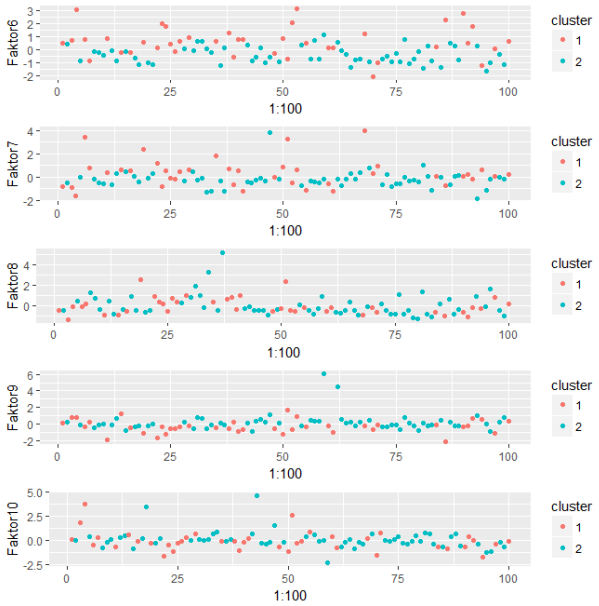


Figure 3. The Result of Laboratory Clustering using *K*-Means for Each Factor

According Figure 3, we can see in plotting factor or variables 1, 3, 4, 6 and 7 with each laboratories, cluster 1 tend to have a higher value than cluster 2, meanwhile in other factors/variables, we cannot draw conclusion because the cluster plotting result is more vague. Take a closer look in factor 1 in Figure 4.

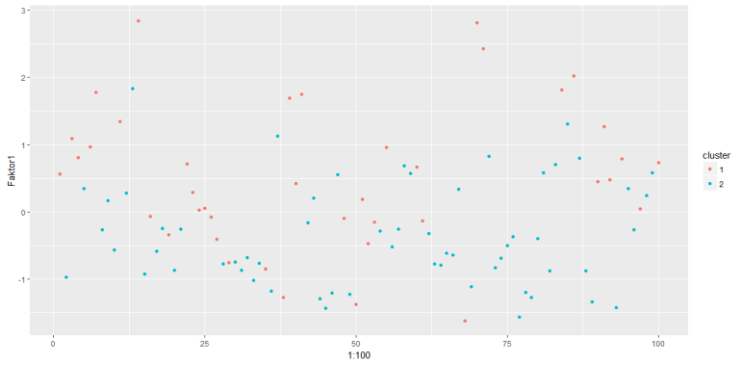


Figure 4. The Result of Laboratory Clustering using *K*-Means for Factor 1

Factor 1 explain about the research productivity of each laboratories. The more productive a lab in their research, they belong to cluster 1 as we can see by the red dot. The less productive lab was clustered in cluster 2.

3.3. *K*-Medoids. *K*-means clustering is has efficient algorithm, especially in large datasets. But, *k*-means is sensitive to outliers. A robust alternative to *k*-means is *k*-medoids, which is based on medoids. The term medoid corresponds to the most centrally located point in the cluster. *K*-medoids algorithm is less sensitive to noise and outliers, compared to *k*-means, because it uses medoids as cluster centers instead of means. The

most common *k*-medoids clustering methods is the PAM (Partitioning Around Medoids) algorithm [5].

To generate *k*-medoids algorithm, first we specify the number of cluster (*k*). One of approach to determine the optimal number of clusters is the silhouette method. Based on the silhouette method, we get *k*=2. The result is same with the purpose of this paper, which to cluster given data set into 2 groups. Figure 5 shows the result of clustering using *k*-medoids with *k*=2.

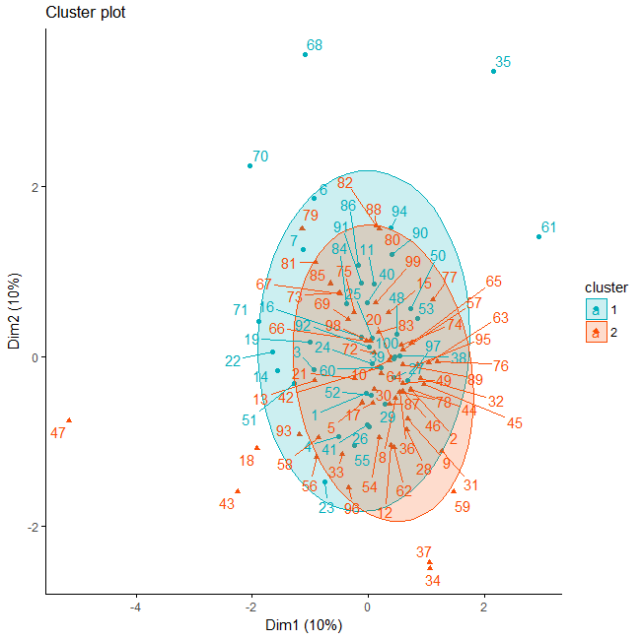


Figure 5. The Result of Laboratory Clustering using *K*-Medoids

There are 40 observations allocated in cluster 1 and 60 observations in cluster 2. The plot above only explain 20% of point variability so we will see cluster position in each factor/variables as in Figure 6.

From Figure 6, we can see that cluster 1 as in red dots tend to have higher value, while cluster 2 as in blue dots tend to have lower value. So from plot above we can tell that cluster 1 contain laboratorium with good or better performance, meanwhile cluster 2 contain laboratorium with poor performance.

 ISSN 2621–086 X

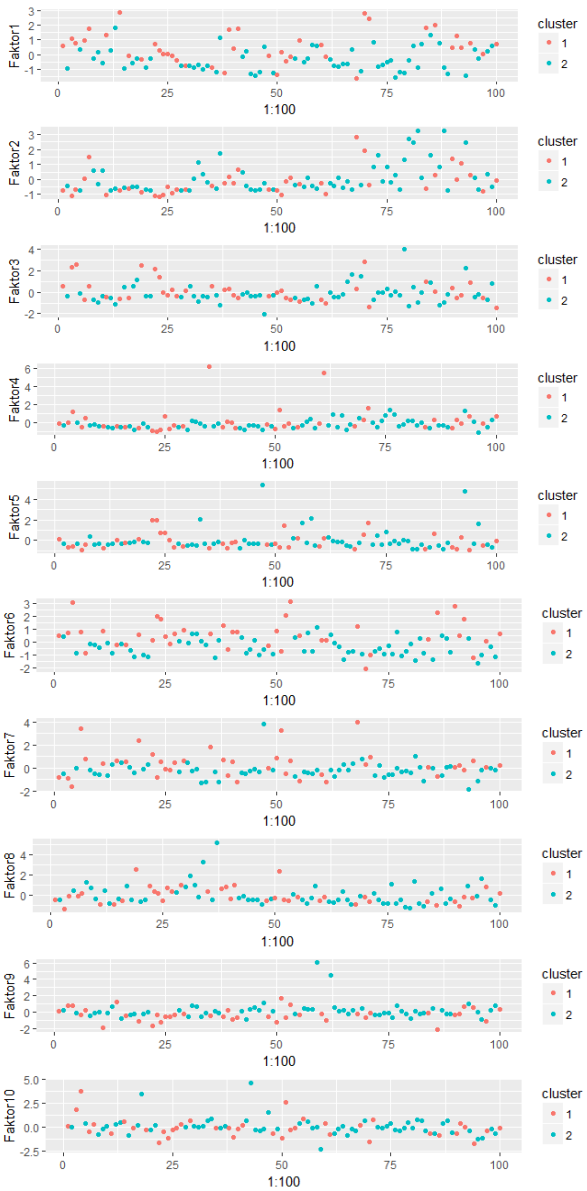


Figure 6. The Result of Laboratory Clustering with *K*-Medoids for Each Factor

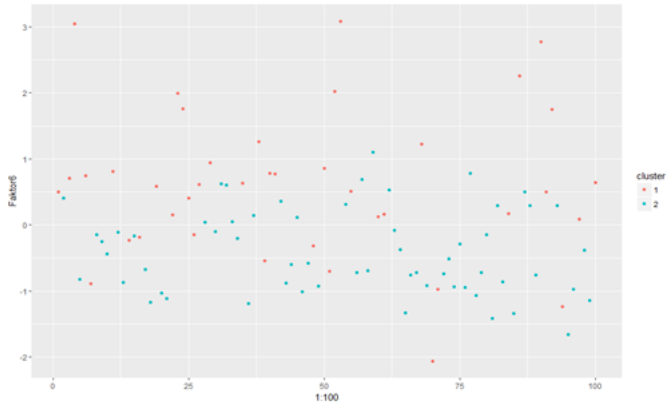


Figure 7. The Result of Laboratory Clustering using *K*-Medoids for Factor 6

Factor 7 explain about laboratories activity in teaching program. From plot above, cluster 1 as seen in red dots has higher value, so it means that laboratories clustered in cluster 1 has more activity in teaching program compare to lab in cluster 2.

3.4. Model-Based Clustering. The standard clustering methods, such as *k*-means clustering and hierarchical clustering, are heuristic and are not based on formal models. Furthermore, *k*-means algorithm is commonly randomly initialized, so different runs of *k*-means will often yield different results. Additionally, *k*-means requires the user to specify the optimal number of clusters, whereas the real cluster is unknown. An alternative is Model-Based Clustering, which consider the data as coming from a distribution that is mixture of two or more clusters [6]. Unlike *k*-means, the Model-Based Clustering uses a soft assignment, where each data point has a probability of belonging to each cluster. For this data, it can be seen that Model-Based Clustering selected a model with three clusters.

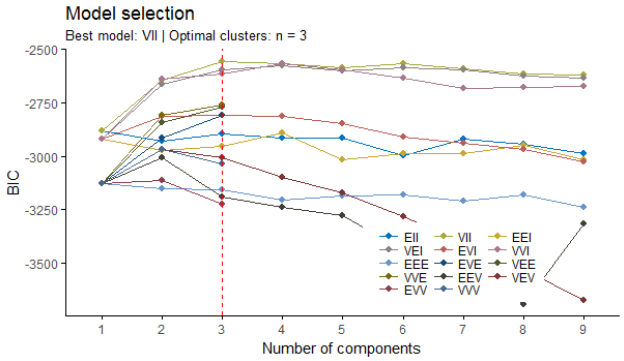


Figure 8. BIC Plot with Vertical Axes Adjusted to Display The Maximum Values

The optimal selected model is VII model which components are spherical with varying volume and equal shape. The visualization data in a scatter plot according to its cluster assignment can be seen in Figure 9.

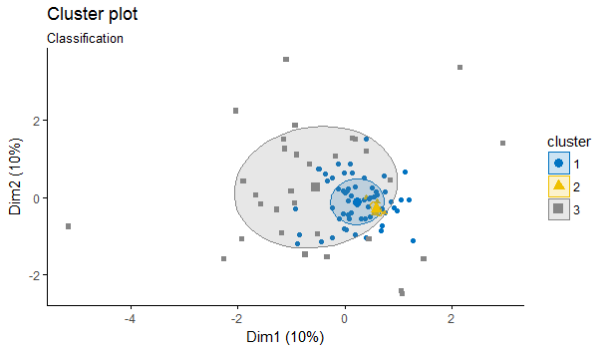


Figure 9. Classification (Left) and Uncertainty (Right) Plots

The blue circle represent lab which was clustered into cluster 1, the orange triangle is lab which was put in cluster 2, and the grey square in cluster 3. The result is 61

laboratories are allocated in cluster 1, 7 laboratories in cluster 2, and 32 laboratories in cluster 3. The plot above only explain 20% of point variability so we will see cluster position in each factor, presented in Figure 9.

From Figure 9, we can see in plotting factor 1, 2, 3, and 7 with each laboratory, cluster 3 tend to have a higher value than the other cluster, so we can know that cluster 3 is laboratories with good performance, meanwhile in other factors, we cannot draw conclusion because the cluster plotting result is vague.

The higher research productivity of each laboratory (factor 1), level of lecture publication (factor 2), international relationship (factor 3), and laboratory activity in teaching program (factor 7) makes the laboratories clustered into cluster 3. The laboratories that not included in cluster 3, especially laboratories that belongs to cluster 2 (poor performance), can improve their performance by focusing on these 4 factors.

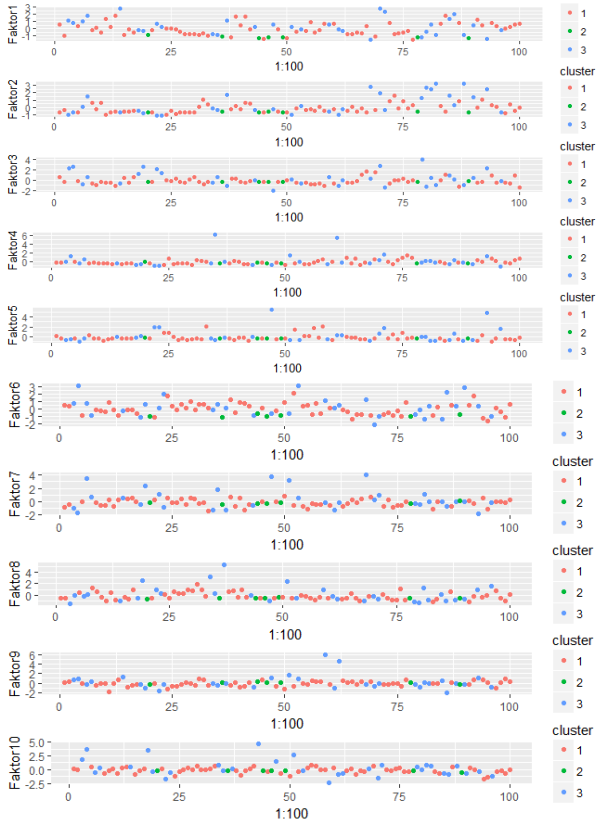


Figure 10. Classification in Each Factor using Model-Based Clustering

This Integrated Convergent Divergent Random (ICDR) is used to evaluate the goodness of the cluster by looking at the value of the distribution of the data in the cluster, the smaller ICDR values indicate the smaller the cluster members' differences in each cluster, which means that each cluster already has members with similar

characteristics. The following is presented comparison of each cluster method using ICDR.

Table 3. The Comparison of ICDR between Cluster Methods

Cluster Method	Number of Cluster	Score
K-Means	2	0.918
K-Medoids	2	0,869
MBC	3	0,113

According to Table 3, we know that the best method based on ICDR is Model-Based Clustering.

4. Conclusion

According to ICDR, the best model for laboratory clustering is Model-Based Clustering. After using Model-Based Clustering, we obtain 7 laboratories with poor performance i.e, Geomarin Laboratory, Interaksi, Grafik, dan Seni Laboratory, Studio Aplikasi Terapan, Studio Sistem Operasi dan Jaringan, Listrik Dasar Laboratory, Teknik Kimia Laboratory, and Fisika Rekayasa. Hence the stakeholder should be focus to improve the performance from these laboratories. Improving the performance of this laboratory can focus on 4 factors like research productivity of each laboratory (factor 1), level of lecture publication (factor 2), international relationship (factor 3), and laboratory activity in teaching program (factor 7).

References

[1] Sonhadji, A. *Laboratorium Sebagai Basis Pendidikan Teknik di Perguruan Tinggi*. Universitas Negeri Malang. Malang. 2002.
[2] Agustini, M. *Model-Based Clustering dengan Distribusi t Multivariat Menggunakan Kriteria Integrated Completed Likelihood dan Minimum Message Length.* Institut Teknologi Sepuluh Nopember. Surabaya. 2017.
[3] Johnson, R. A. and Dean W. W. *Applied Multivariate Statistical Analysis Sixth Edition*. Person Prentice Hall. New Jersey. 2007.
[4] Banfield, J. D. and Raftery, A. E. Model-Based Gaussian and non-Gaussian Clustering. *Biometrics.* 49(3): 803-821. 1993.
[5] Park, H. S. and Jun, C. H. A Simple and Fast Algorithm for K-Medoids Clustering. *Elsevier*. South Korea. 2008.
[6] Fraley, C. and Raftery, A. E. Model-Based Clustering, Discriminant Analysis, and Density Estimation. *Journal of the American Statistical Association*. 97(458): 611–631. 2002.