Jurnal Bayesian : Jurnal Ilmiah Statistika dan Ekonometrika
p-ISSN: 2775-7463 | e-ISSN: 2775-7455 Vol.3 No. 1 Maret 2023
Doi Issue: doi.org/10.46306/bay.v3i1 Doi Article: doi.org/10.46306/bay.v3i1.52

COMPARISON PERFORMANCE NON-HIERARCHICAL CLUSTER (Case Study: Central Java Regional Competitiveness Index, 2022)

Pardomuan Robinson Sihombing[1], Ade Marsinta Arsani[2], Dyah Purwanti[3], Sigit Budiantono[4]
[1,2]Badan Pusat Statistik Indonesia
[3,4]PKN STAN
Email: *robinson@bps.go.id*

ABSTRACT: The Regional Competitiveness Index (RCI) is a benchmark for measuring a region's ability to compete in a market. RCI covers several indicators, including infrastructure, human resource quality, innovation, and government policies supporting economic growth. This study aims to test the performance of several non-hierarchical cluster techniques. The data used Regional Competitiveness Index data in 35 Cities in Central Java in 2022 from the National Research and Innovation Agency (BRIN). The optimal number of clusters recommended using the Elbow method technique is as many as 3. The K-Means method is the best considering the largest Silouhette and R2 values and the smallest AIC/BIC. Cluster 1 has negative values for Pillars 2, 4, 9, and 10. Members in this cluster are Sukoharjo, Magelang City, Surakarta, Salatiga, Pekalongan City, and Tegal City. On the other hand, Cluster 2 has only one negative value for pillar nine. The members of this cluster are Semarang City. The third cluster is only positive in pillar nine and pillar 28. The members of this cluster are as many as 28 other districts. A comprehensive and targeted policy is needed so that the competitiveness index of the Central Java region continues to increase.
Keywords: Cluster, Fuzzy C-Means, K-Means, K-Median, K-Medoid, RC

INTRODUCTION

The Regional Competitiveness Index (RCI) is a benchmark used to measure the ability of a region to compete in a market. RCI includes several indicators, such as infrastructure, quality of human resources, innovation, and government policies that support economic growth. One of the region whose RCI value continues to increase in Central Java. The development of RCI of Central Java Province has increased in line with the government's efforts to improve infrastructure, improve the quality of human resources, and encourage innovation and investment in the area. In 2022, RCI Central Java Province scored 3.63 and ranked 4th out of all provinces in Indonesia.

To continue to improve RCI, the government conducts various efforts and policies. Of course, the actions and procedures carried out must remain targeted by looking at each condition of a region. Therefore, it is necessary to group regions based on the RCI value of each pillar so that the policies carried out are to the needs of each region. One of the analyses in statistics used to group subjects is cluster analysis. The multivariate analysis includes cluster analysis (Rencher &; Christensen, 2012). In general, cluster analysis can be divided into hierarchical and non-hierarchical clusters (Johnson & Dean, 2008). Hierarchical clusters group data based on distance and correlation between variables.

Meanwhile, in non-hierarchical clusters, researchers have determined the desired number of groups. Non-hierarchical clusters are sometimes considered more efficient than hierarchical clusters, especially for large numbers of observations. In addition, non-hierarchical clusters are considered more accessible to interpret the results of their analysis. One method in a non-hierarchical cluster is k-means, k-median, k-medoid and fuzzy k-means.

Each of the methods has its advantages and disadvantages. The k-mean clustering method looks for cluster centres calculated based on the average distance between data. This method is faster and more suitable for big data. The k-median clustering method looks for cluster centres calculated based on the median distance between data. This method is more stable against extreme data. K-medoid method: The k-medoid clustering method looks for cluster centres calculated based on data objects in that cluster. This method is more robust against data that is not symmetrical or with extreme value.

Jain and Murty (1999) compare clustering methods, including K-Means and K-Medoid. The author suggests that K-Medoid is more suitable for use on data with extreme values or not symmetrical data because this method is more robust to the data than K-Means. However, K-Medoid requires longer computation time compared to K-Means.

Yin et al. compare k-means and k-median. The author states that k-means is better for homogeneous data, while k-median is suitable for data containing outlier data (Yin et al., 2013) Park and Jun (2009) compared the K-Medoid clustering method with Fuzzy C-Means using data from sea level and aerial observation data. The results showed that the K-Medoid clustering method is better at determining stable cluster centres and is not sensitive to initialization.

Yu and Wang (2011) compared the K-Means clustering method with Fuzzy C-Means on medical data. The results showed that the Fuzzy C-Means method performed better than the K-Means in determining the right cluster for medical data.

Based on the above problems, there is still a gap in research results between the K-Means, K-Median, K-Medoid and Fuzzy C-Means methods. Researchers are interested in testing the performance of the fourth K non-hierarchical cluster method in grouping city districts in Central Java based on DSD pillar I data in 2022.

METHODS

The data used in this study comes from the publication of the Research and Innovation Agency (BRIN, 2023). The study used district-level data from cities in Central Java. This research only uses data on 11 pillars in RCI from 12 existing pillars. This data is because pillar 11 has the same value for all observations; the comments are worth 5 for 35 city districts. Because all variables have the same unit, there is no need to transform data using logarithms or standard data values (*z score*).

K-Means Cluster

K-Means is one of the multivariate analyses and one of the methods in data mining with *unsupervised* techniques that group data with a *partitional* system. This method works by collecting data in one cluster based on the similarity or proximity of characteristics to other cluster data. The K-Means method will be general.

K-Median Cluster

K-Median cluster is one of the clustering techniques used to group data into groups or sets (Han et al., 2011). This method finds the K-Means based on the median distance between data. The goal is to minimize the distance between the K-Means and the data contained in the cluster.

K-Medoid

K-Medoid is a clustering method used to partition data into clusters based on the distance of data objects to the cluster's centre. The cluster centre in the K-Medoid method is determined by one of the data objects in the cluster, while the other data objects are calculated by the distance to the cluster centre (Kaufman &; Rousseeuw, 1990). The K-Medoid method is very suitable for use on data with extreme values or data that is not symmetrical, because this method is more robust to the data than other clustering methods such as K-Mean. However, the disadvantage of the K-Medoid method is the relatively longer computational time than the K-Mean.

Fuzzy C-means (FCM)

Fuzzy C-Mean is a development of K-Means by combining *fuzzy* principles with the K-Means method. The difference is that data clustered using FCM will be a member of each *existing cluster*. Data binding to *a cluster* is determined by its membership value, which ranges from 0 to 1.

Elbow Method

The elbow method is a method that is often used to determine the number of clusters to be used in k-means clustering by looking at the percentage of comparison results between the number of clusters that will form elbows at a point (Madhulatha, 2012). In general, the results of different percentages of each cluster value can be shown using graphs as a source of information.

Model Selection Criteria

In this study, the model selection was based on silhouette criteria (Struyf et al., 1997), where the model with the greatest value was chosen. As for the formula used:

$$s(i) = \frac{b(i)-a(i)}{max\{a(i)-b(i)\}} \quad (1)$$

with:

$$a(i) := \frac{\sum_{j \epsilon A, j \neq 1} d(i,j)}{|A|^{-1}} \quad (2)$$

$$d(i) := \frac{\sum_{j \epsilon C,} d(i,j)}{|C|} \quad (3)$$

$$b(i) := \min_{C \neq A} d(i,j) \quad (4)$$

where

A = amount of data in cluster A

d(i)= distance

b(i) =minimum value of the average distance of i-th data with all data in different clusters.

C = amount of data in cluster C

In addition, error criteria are used, including AIC (Akaike, 1974) and BIC (Gideon Schwarz, 1978) and coefficient of determination (R^2). The best model is the model that

has the smallest AIC and BIC values (Widarjono, 2007) and the most significant coefficient of determination (Gujarati, 2004). The formula used is:

$$AIC = -2\,L(\hat{\theta}) + 2p \quad (5)$$

$$BIC = -2\,L(\hat{\theta}) + p\ln(n) \quad (6)$$

$$R^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \quad (7)$$

Where is the likelihood value, p is the number of parameters to be estimated, including constants, and n is the number of samples. Value is the predicted value of the model's dependent variable, and Y is the observation value of the dependent variable.$L(\hat{\theta})\hat{Y}$

RESULTS AND DISCUSSION

Before further discussing the grouping of provinces, a descriptive analysis was carried out on the research variables. Table 1 presents descriptive statistics of each RCI pillar. The pillar scored highest on the 11th pillar of business dynamism, followed by the institutional and Health pillars. The pillar with the highest data diversity is the pillar of market growth.

Table 1. Descriptive Statistics of Research Variables1

Descriptive Statistics	Mean	Std. Deviation	Minimum	Maximum
Pillar 1: Institutions	4.449	0.128	4.17	4.62
Pillar 2: Infrastructure	2.149	0.487	1.53	3.3
Pillar 3: ICT Adoption	3.303	0.348	3.05	4.95
Pillar 4: Macroeconomic stability	2.978	0.347	2.35	4.17
Pillar 5: Health	4.241	0.138	3.85	4.46
Pillar 6: Skills	3.066	0.468	2.4	4.14
Pillar 7: Product market	2.923	0.725	1.06	4.33
Pillar 8: Labor market	2.995	0.428	2.42	4.04
Pillar 9: Financial system	3.364	0.743	2.11	5
Pillar 10: Market size	1.709	1.119	0.43	5
Pillar 11: Business dynamism	5	0	5	5
Pillar 12: Innovation Capabilities	1.975	1.111	0.97	4.99

In Figure 1 can be seen the number of optimal clusters recommended by the Elbow method with the smallest BIC criteria. Using the Elbow Plot method, 3 clusters were selected in the K-Mean s and K-Median methods and 2 clusters for K-Medoid and Fuzzy Means.

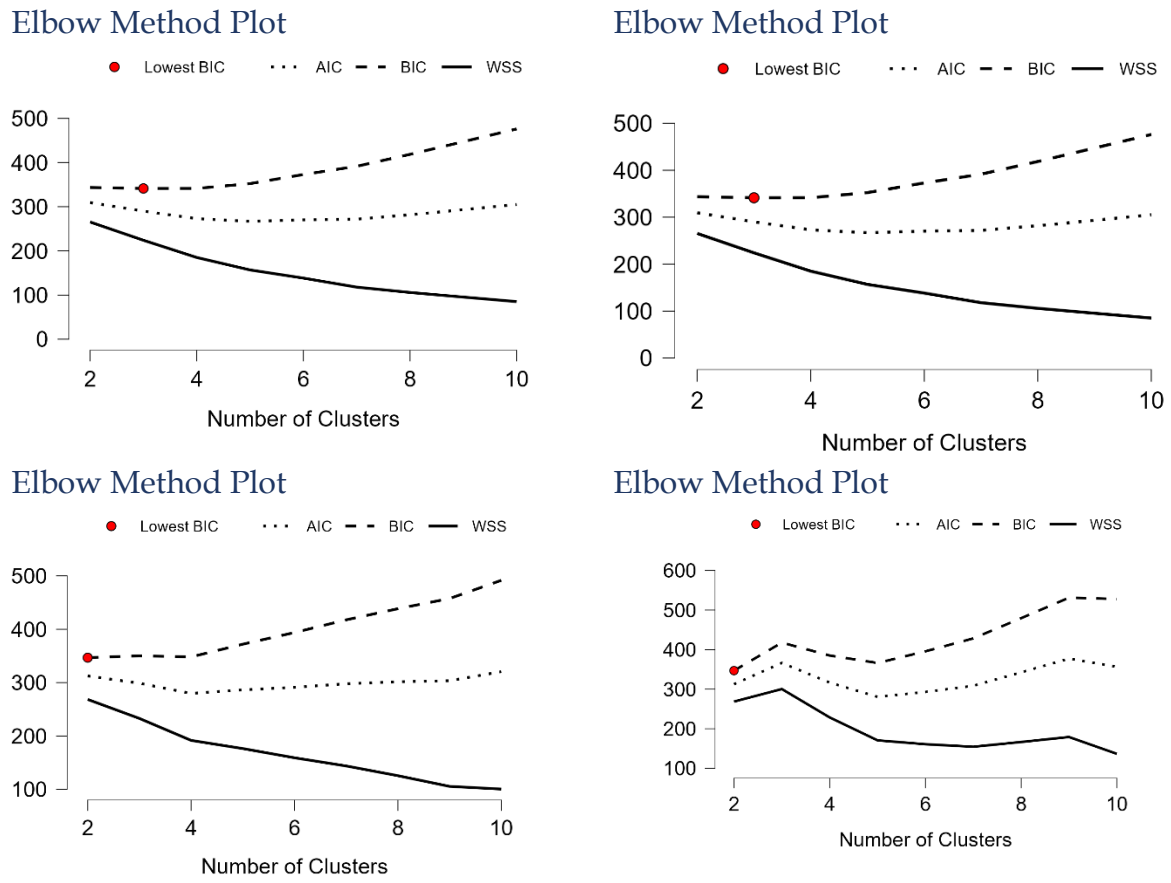

Figure 1. Selection of Multiple Clusters in K-Means, 1K-Median, K-Medoid and Fuzzy Means with Elbow Plot

Furthermore, the best model was selected by comparing each method's criteria: the Silhouette index, error criteria (AIC and BIC) and the coefficient of determination R^2. When viewed from the most significant Silhouette and R2 index values, the best method is K-Means. Meanwhile, if assessed from the AIC and BIC values, the K-Means method has smaller values, namely 290.03 and 341.15. So it can be said that the K-Means model is better than the three non-hierarchical cluster models in grouping city districts in Central Java based on RCI values because of the smaller AIC and BIC values and the larger R^2 and silhouette.

Table 2. Best Model Selection Criteria

Clustering	Sum	R^2	AIC	BIC	Silhouette
K-Means	3	0.401	290.03	341.35	0.30
K-Medians	3	0.387	299.06	350.39	0.14
K-MedoRCI	2	0.378	312.53	346.74	0.34
Fuzzy C-Means	2	0.100	363.94	398.16	0.08

The K-Means clustering method has several advantages compared to other clustering methods, including relatively faster computational acceleration than other clustering methods such as K-Medoid or Hierarchical Clustering. In addition, this method has been implemented and is easy to understand so that it can be used by beginners in the

field of clustering. On the other hand, the k-means shake method is used on data with a spherical shape and large dimensions (A. K. Jain &; Ambassador, 1988).
Table 3 shows the average value of each variable per cluster. A positive value indicates that the value of the group variable in the cluster is below the average of the overall data. In contrast, a negative value indicates that the group variable's weight exceeds the data's general standard. Cluster 1 has negative values for pillar 2, pillar 4, pillar 9 and pillar 10. Members in this cluster are Sukoharjo, Magelang City, Surakarta, Salatiga, Pekalongan City, and Tegal City. This result indicates that the six city districts are still lagging in infrastructure, economic stability, financial system and market stability.

Table 3. Cluster Means Each Variable

Cluster Means	Cluster 1	Cluster 2	Cluster 3
Pillar 1: Institutions	0.692	0.316	-0.16
Pillar 2: Infrastructure	-0.29	2.364	-0.022
Pillar 3: ICT Adoption	0.865	4.74	-0.355
Pillar 4: Macroeconomic stability	-0.038	3.431	-0.114
Pillar 5: Health	0.647	1.588	-0.195
Pillar 6: Skills	1.353	1.865	-0.356
Pillar 7: Product market	1.276	1.444	-0.325
Pillar 8: Labor market	1.541	1.389	-0.38
Pillar 9: Financial system	-0.057	-1.687	0.072
Pillar 10: Market size	-0.559	2.941	0.015
Pillar 12: Innovation Capabilities	1.46	2.714	-0.41

On the other hand, Cluster 2 has only one negative value for pillar 9. The members of this cluster are Semarang City. The third cluster is only positive in Pillars 9 and Pillars 10. The members of this cluster are as many as 28 other districts.

Table 4. Members of each cluster

Cluster	Member
Cluster 1	Sukoharjo, Magelang City, Surakarta, Salatiga, Pekalongan City, Tegal City
Cluster 2	Semarang City
Cluster 3	Banyumas, Purbalingga, Banjarnegara, Kebumen, Purworejo, Wonosobo, Magelang District, Boyolali, Klaten, Wonogiri, Karanganyar, Sragen, Grobogan, Blora, Rembang ,Pati , Kudus, Jepara, Demak, Semarang, Temanggung, Kendal, Batang, Pekalongan, Pemalang, Tegal, Brebes

CONCLUSION

The number of optimal clusters aligned using the Elbow method technique with BIC criteria is 3 clusters in the K-Means and K-Median methods and 2 clusters for K-Medoid and Fuzzy C-Means. Considering the largest Silouhette and R^2 values, the K-Means method is the best among the four methods. This method is also supported based on the more efficient AIC and BIC values criteria.

For further research, you can add other potential variables such as the Happiness Index, Youth Development Index, Cultural Index, Employment Index, Village Development Index, and Building Village Index. In terms of methods, you can add neighbourhood methods, random forest methods and cluster hierarchy methods.

REFERENCES

Akaike, H. (1974). A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

BRIN. (2023). *Indeks Daya Saing Daerah 2022*.

Gideon Schwarz. (1978). Estimating The Dimension of a Model. *The Annals of Statistics*, *6*(2), 461–464.

Gujarati, D. (2004). *Basic Econometrics BY Gujarati* (pp. 1–1002). McGraw-Hill Inc.

Han, J., Kamber, M., & Pei, J. (2011). *Data mining: concepts and techniques*. Morgan Kaufmann.

Jain, A. K., & Dubes, R. C. (1988). *Algorithms for clustering data*. Prentice-Hall, Inc.

Jain, A., & Murty, M. (1999). Data clustering: A review. *ACM Computing Surveys (CSUR)*, *31*(2), 264–323.

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons.

Madhulatha, T. . (2012). An Overview On Clustering Methods. *IOSR Journal Engineering*, *2*(4), 719–725.

Park, H. S., & Jun, C. H. (2009). A simple and fast algorithm for K-medoids fuzzy clustering. *Expert Systems with Applications*, *36*(2), 3336–3341.

Struyf, A., Hubert, M., & P. J. Rousseeuw. (1997). Clustering in an Object-Oriented Environment. *Journal of Statictical Software*, *1*(4), 1–30.

Widarjono, A. (2007). *Ekonometrika: Teori dan Aplikasi untuk Ekonomi dan Bisnis*. Ekonosia Fakultas Ekonomi Universitas Islam Indonesia.

Yin, J., Li, H., & Shen, J. (2013). Comparison of K-Means and K-Medians Clustering on Different Datasets. *In Applied Mechanics and Materials*, *295*, 760–763.

Yu, L., & Wang, S. (2011). Comparison of K-means and fuzzy C-means clustering on a medical data set. *Journal of Medical Systems*, *35*(5), 703–710.