Parameter: Journal of Statistics
Available at https://bestjournal.untad.ac.id/index.php/parameter
Vol. 2 No. 2 2022, 18 - 25
DOI: *https://doi.org/10.22487/27765660.2022.v2.i2.15898*
eISSN: 2776-5660

FACTOR ANALYSIS FOR INCREASING READING LITERACY IN INDONESIA

Rizka Pitri[1*], Ayu Sofia[2]
[1]Raden Intan State Islamic University, Indonesia
[2] Sumatera Institute of Technology, Indonesia

[*]***e-mail:*** *rizka@radenintan.ac.id*

ABSTRACT

Low interest in reading is a problem for our nation that must be solved, because Indonesia to occupy second position from bottom in terms of literacy. Most of the provinces in Indonesia are at low literacy activity levels and none of the provinces are included in the high literacy activity level. The lack of interest in reading can be influenced by many factors. Access of supporting resource where people get literacy materials, such as libraries, bookstores, and mass media, how people to get the information technology, and media devices to access literacy materials are the factor that can be affect the interest of reading. Literacy is one of the important cultures for a country. That is because the culture is able to influence the intelligence and well-being of a country's life. So the study aims to see what factors affect to increasing the literacy reading in the provinces in Indonesia. This study uses k-means clustering before applying factor analysis. Based on k-means clustering, two clusters are formed and showed one of the cluster showed that the second cluster is the provinces that have the highest number of library's facilities. In addition based on the analysis factor in each cluster, two factors were formed, namely the standard factor for reading literacy levels and supporting the facilities for reading literacy. It can be concluded that the way to increase reading literacy in two clusters of the area in Indonesia are by increasing the standard of reading literacy level and supporting the facilities for reading literacy.

Keywords: *Factor Analysis, K-means Clustering, Reading Literacy*

Cite: Pitri, R., & Sofia, A. (2022), Factor Analysis for Increasing Reading Literacy in Indonesia. *Parameter: Journal of Statistics*, *2*(2), 18-25. https://doi.org/10.22487/27765660.2022.v2.i2.15898

[*]Received: May 2022; Reviewed: May 2022; Published: June 2022

INTRODUCTION

Reading can be one way to get information. Reading can also be interpreted as a thought in a process of reasoning. By reading we involve the process of thinking to understand the content of the media we read. Reading has a very vital role in contributing to the golden generations of carriers of progress, of course we agree that reading will increase intelligence and knowledge. In addition, reading can also cause many benefits such as improving brain performance, increasing knowledge, and sharpening memory.

UNESCO considers Indonesia's reading interest to be very concerning, namely with a percentage of 0.001%. So, out of 1000 people in Indonesia there is only one person who diligently reads. This caused Indonesia to occupy the bottom second position in terms of literacy. In addition, the study of the Program for International Student Assessment (PISA) released by the Organization for Economic Co-Operation and Development (OECD) in 2015, stated that Indonesia was in the position of 62 out of 70 countries studied in terms of literacy with an Indonesian reading score of 397, lower than the average score of 493. World's Most Literate Nation Ranked research conducted by Central Connecticut State University (CCSU) in 2016 showed that Indonesia was ranked 60th out of 61 countries for reading interest. This is a big problem in the field of Indonesian education that will have an impact on the future development of the nation.

Low interest in reading is a problem for our nation that must be solved, because the lack of interest in reading can be influenced by many factors. Sutarno (2006) grouped various factors that affect reading interest, one of which is the availability of interesting, quality, and diverse reading materials in the community. Miller and McKenna (2016) raised the following factors that affect literacy activities, namely: (1) Proficiency is an initial requirement so that a person can access literacy materials; (2) Access is a supporting resource where people get literacy materials, such as libraries, bookstores, and mass media; (3) Alternatives are a wide selection of information technology and entertainment devices to access literacy materials; and (4) Culture is a habit that also forms the habitus of reading literacy.

The library is a medium in distributing reading materials. The availability of libraries is a determinant that affects people's reading interest because not all regions have decent, complete and easy-to-reach libraries. But accompanied by the development of technology, the development of the internet and the development of new sources of information so fast that it requires libraries to make a step change, both in the form of collections and in terms of service patterns. The library also comes in digital form so that it can be accessed anytime and anywhere. The presence of digital libraries can be an alternative for the community because users are no longer physically attached to library service hours where users must attend or visit the library to get information. So that this becomes very closely related to the availability of reading materials both in physical and digital form accompanied by technological advances in internet access as a determining factor for low interest in reading by the community.

The results of the Provincial Reading Literacy Activity Index were conducted by the Research Team of the Center for Education and Culture Policy Research (Puslitjakdikbud), Balitbang Kemendikbud in 2018. Of the thirty-four provinces in Indonesia, 9 provinces (26%) fall into the category of moderate literacy activities; 24 provinces (71%) fall into the low category; and 1 province (3%) is in the very low category. This means that most provinces are at low literacy activity levels and none of the provinces are included in the high literacy activity level. So that the study aims to see what factors affect to increasing reading literacy in the provinces in Indonesia that will be divided into several clusters.

MATERIALS AND METHODS

1. Materials

The research data source is based on the results of the National Socioeconomic Survey (SUSENAS) in 2020 and Perpusnas's data in each province of Indonesia. There are 8 variables used in this study, namely:

a. Community literacy development index (IPLM) as x_1 (Perpusnas's data)
b. Number of districts receiving mobile library cars as x_2 (Perpusnas's data)
c. Number of Libraries as x_3 (Perpusnas's data)
d. Number of Book Collections as x_4 (Perpusnas's data)
e. Reading Enjoyment Level (TGM) as x_5 (Perpusnas's data)
f. Number of library digital applications as x_6 (Perpusnas's data)
g. Literacy Rate as x_7 (SUSENAS)

h. Percentage of people using internet as x_8 (SUSENAS)

2. Methods

a. K-means Clustering

Clustering is the process of grouping a set of data into clusters that have similarities. One of the methods used to find clusters in data is K-Means clustering where k presents the number of clusters [1]. A data is grouped into one cluster based on the similarity of attributes owned. This similarity can be known by applying distance measurements. The method of calculating distance in this study is Euclidean Distance. Euclidean distance is the most commonly used distance calculation. For 2 x and y data points in the d-dimension of the data, the calculation of distance using Euclidean distance is formulated with Equation (1).

$$d_{euc}(x,y) = \sqrt{\sum_{j=1}^{d}(x_j - y_j)^2} \quad (1)$$

Where: x_j, y_j = value of j attribute

b. Silhouette Coefficient

To see the quality of the results of the cluster of each distance calculation, it is necessary to conduct a homogeneity test. The test is performed after reaching convergence 0 where the result of the last cluster is the same as the previous cluster. In other words, no data moves clusters. The test is calculated using the Silhouette coefficient equation. The step in calculating the Silhouette coefficient begins by finding the average distance of the ith data with all the data in the same cluster, here we assume the-ith data is in cluster A. The formula of a(i) is written in Equation (2).

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

Where: A= total data in cluster A

Next calculate the value b(i) which is the minimum value of the average distance of the ith data with all the data in the cluster is different. Now, let's assume the cluster is different other than A with cluster C. Then, the calculation of the average distance of the ith data with all data in cluster C is written as follows:

$$d(i,C) = \frac{1}{|C|}\sum_{j\in C} d(i,j) \quad (3)$$

where C = total cluster di C

After calculating d(i,C) for all clusters C ≠ A, then select the minimum distance value as the value b(i).

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

If cluster B has a minimum distance value, then d(i,B) = b(i) which is referred to as the neighbor of the ith data and is the second best cluster for the ith data after cluster A. After a(i) and b(i) are known, then the final process calculates silhouette coefficient [16]

$$s(i) = \frac{b(i)-a(i)}{\max\{a(i)-b(i)\}} \quad (5)$$

Value $s(i)$ in range -1 until 1, where each value is interpreted as follows:
$s(i) \approx 1$ then the ith data is well trusted (in A)
$s(i) \approx 0$ then the ith data is in the middle between the two clusters (A and B)
$s(i) \approx -1$ then the ith data weakly classed (close to klister B rather than A)

Table 1. Silhouette coefficient

Silhouette coefficient	Interpretation
0.71 – 1.00	The result of the structure is strong
0.51 – 0.70	The result of the structure is good
0.26 – 0.50	The result of the structure is weak
≤ 0.25	Unstructured

c. Factor Analysis

Factor analysis is a multivariate analysis technique used to reduce data or summarize from variables that are widely converted into little so-called factors and still contain most of the information contained in the original variable. Before the factor analysis is carried out, an examination of the relationship between the changer of each group and the KMO test and the Bartlett test.

1. KMO (Kaiser-Meyer-Olkin Test)

The Kaiser-Meyer-Olkin (KMO) test is needed to see the adequacy of the analyzed sample (sampling adequacy). This KMO value is obtained by comparing the magnitude of the observed correlation coefficient with the magnitude of the partial correlation coefficient. Where the size of KMO with a value of <0.5 is considered less suitable for the variable. For more details, the following is the KMO formula:

$$KMO = \frac{\sum_{i=1}^{p}\sum_{j=1}^{p} r_{ij}^2}{\sum_{i=1}^{p}\sum_{j=1}^{p} r_{ij}^2+\sum_{i=1}^{p}\sum_{j=1}^{p} a_{ij}^2} \quad (6)$$

Where:

$i = 1,2,3,..,p$ =dan $j = 1,2,3,..p$

r_{ij} = observed correlation coefficient between variables I and j

a_{ij} = partial coefficient between variables I and j

2. Bartlett's Test

The Bartlett test aims to find out if there is a relationship between variables. If the variables X1, X2,...,Xp are independent, then the correlation matrix between variables is the same as the identity matrix. The Bartlett test has a high accuracy (significance) p < 0.00000, giving the implication that the correlation matrix is suitable for factor analysis. Bartlett's test results are the result of a test on the hypothesis:

Ho : Correlation matrix = identity matrix
Hi : Correlation matrix ≠ identity matrix

Rejection of Ho is done by comparing Bartlett test value > chi-square table or with significance value < significance level of 5%. If Ho is rejected then the analysis deserves to be used in factor analysis. The Bartlett test is formulated as follows:

$$bartlett\ test = -\ln|R|\left[n - 1 - \frac{2p+5}{6}\right] \quad (7)$$

Where:

$|R|$= determinant value

n= number of data

P= number of variable items

Hair and Anderson (1998) stated that there are several criteria in determining a number of factors formed, one of which is the criteria of eigen value. The reason for using the eigenvalue is because each variable has a value contribution of 1 to the total value of the eigen. So that factors with eigen value ≥ 1 which is considered significant, while for factor <1 is considered insignificant.

The next step in factor analysis is the rotation of factors, namely looking for factors that are able to optimize correlations between observed indicators. Rotation of this factor is necessary if the factor extraction method has not been able to produce a clear main factor component. An overview of the purpose of rotating factors is to be able to obtain a simpler factor structure so that it is easy to interpret. Varimax rotation is the rotation that makes the number of variants of the factors containing the loading square in each factor to the maximum (Johnson and Wichern, 2002). This rotation method seeks to maximize the saboteur factor and result in the origin variable will only have a high and strong correlation with certain factors only (the correlation is close to 1) and have a weak correlation with other factors (the correlation is close to 0).

RESULTS AND DISCUSSION

People's interest in reading in Indonesia still needs to be improved, one of which is to increase the role of libraries. The government's role in increasing the reading interest of the Indonesian is to establish and facilitate the various book collections in each province. Although the government has supported the facilities, for example the libraries, the reading literacy in Indonesia is still low. Figure 1 is a distribution of Community literacy development index (IPLM) score. That figure shows the gradation color that have meaning if an area has a color is more dark, so it means that area has higher IPLM score than other area. But, if the opposite, so the area has lower IPLM score than other area. Figure 1 shows that there are still many provinces that have a "Community literacy development index (IPLM)" under 30. The South Kalimantan is the province that gets the highest IPLM, 48.7.

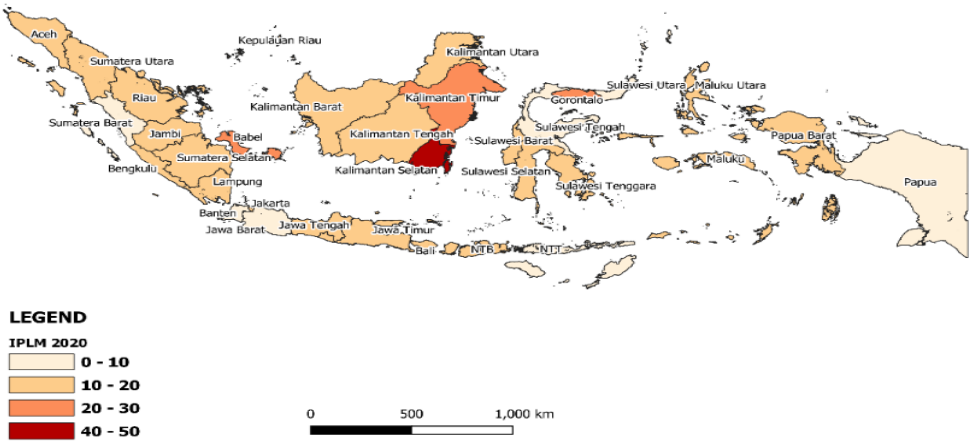

Figure 1. Distribution of Community literacy development index (IPLM) 2020

The dynamics of reading literacy growth in each province is various depending on the awareness of the community and the facilities that support the increasing of reading literacy. Based on the K-means clustering, two clusters were formed. This is based on the highest silhouette value obtained (Table 2).

Table 2. Silhouette Value of each Cluster

	Cluster				
	2	3	4	5	6
Silhouette Value	0.8059	0.6881	0.6255	0.6130	0.5870

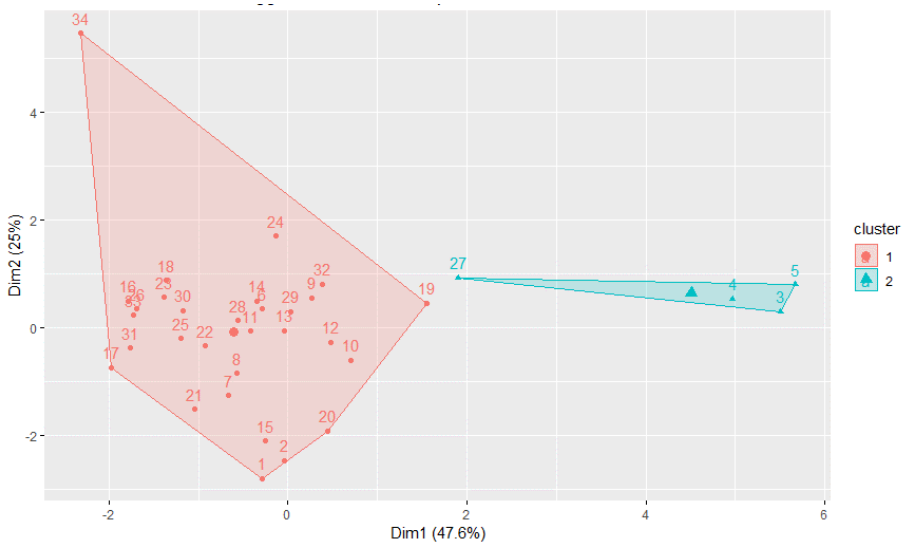

Figure 2. Visualization of K-means Clustering

Based on the k-means clustering, the member of 2nd cluster is the province that has the largest number of libraries, book collections, and districts receiving the mobile library cars in Indonesia. So it can be concluded that the four provinces have adequate supporting facilities for increasing the reading literacy than other provinces. The member of each cluster are showed in Table 3.

Table 3. Member of Each Cluster

Name of Cluster	**Province**
1st cluster	Jakarta, Yogya, NTB, Bangka Belitung, Bali, West Sumatera, Banten, Jambi, Lampung, Riau, South Sumatera, South Kalimantan, West Sulawesi, Gorontalo, Central Sulawesi, South Sulawesi, East Kalimantan, Kepulauan Riau, Central Kalimantan, North Sulawesi, NTT, Bengkulu, North Maluku, Southeast Sulawesi, West Kalimantan, Maluku, North Kalimantan, Aceh, West Papua, Papua
2nd cluster	West Java, East Java, Central Java, North Sumatera

After clustering analysis, factor analysis was carried out. Factor analysis was used to summarize all the dependent and interdependent variables. Factor analysis was applied in each cluster. So the formed factors can represent the variables that are the members of these factors. Before applying the factor analysis, do the checking of relationship between the variables in each cluster. This checking is seen based on the results of the Bartlett test of sphericity and the KMO measure of sampling adequacy. The result of the Bartlett test of sphericity and the KMO measure of sampling adequacy are showed in Table 4.

Table 4. Score of KMO and P-value of Bartlett Test

Type of Test	**1st Cluster**	**2nd Cluster**
Score of KMO measure of sampling adequancy	0.7	0.5
P-value of Bartlett test of sphericity	1.016e-07	0.0000

Based on Table 4, can be concluded that there is a relationship between IPLM, Reading Enjoyment Level (TGM), the number of districts receiving mobile library cars, the number of book collections, the number of libraries, the number of library digital applications, literacy rates, the percentage of people using internet in each cluster. This can be seen for the p-value generated by the Bartlett test. In addition, based on the results of the KMO measure of sampling adequacy in each cluster, the KMO score is above 0.5, meaning that the number of samples used in this factor analysis is representative enough.

After doing the Bartlett test, then determine the optimal number of factors to use factor analysis. Determination of the optimal number of factors is based on the eigenvalue that must exceeding one. Based on the eigenvalues, each cluster will create two optimal factors. The eigenvalues of each factor

are showed in Table 5. The cumulative variance can be explained by using two factors, namely 64% of diversity the independent variables used in this factor analysis.

Table 5. Eigenvalues of Each Cluster

Type of Cluster	**Eigen Value**							
	1st	2nd	3rd	4th	5th	6th	7th	8th
1st cluster	2.701	2.398	0.892	0.692	0.501	0.380	0.248	0.184
2nd cluster	5.697	1.535	0.766	4.33×10^{-16}	1.33×10^{-16}	5.03×10^{-17}	-6.4×10^{-17}	-1.6×10^{-16}

Factor analysis produces a factor matrix that obtained of coefficient values in each independent variables that have been standardized and expressed as a factor. The coefficient value is namely of the loading value. The loading value is a value that represents the relationship between factors and independent variables. A high loading value means that the independent variables has a high contribution in the factors or in other word the independent variable will be a member of the factor. Table 6 is a factor matrix after having rotation using varimax rotation in each cluster. This step is done, so the loading value can be seen clearly the difference between one factor and another, and makes it easier to cluster the members.

Table 6. Factor Matrix Using Varimax Rotation in Each Cluster

Type of Cluster	**Independent Variable**	**1st Factor**	**2nd Factor**
1st Cluster	IPLM	**0.48**	-0.13
	Number of districts receiving mobile library cars	-0.53	**0.70**
	Number of Book Collections	0.52	**0.65**
	Number of Libraries	0.01	**0.93**
	TGM	**0.75**	0.37
	Number of Library digital applications	0.00	**0.77**
	Literacy Rate	**0.69**	0.05
	Percentage of people using internet	**0.86**	-0.09
2nd Cluster	IPLM	-0.87	**0.47**
	Number of districts receiving mobile library cars	0.09	**0.98**
	Number of Book Collections	**0.9**	0.13
	Number of Libraries	**0.92**	0.27
	TGM	**0.80**	0.55
	Number of Library digital applications	**0.70**	0.34
	Literacy Rate	**0.99**	0.14
	Percentage of people using internet	**1.00**	-0.04

In 1st cluster for 1st factor, it can be seen that the highest loading values are IPLM (0.48), TGM (0.75), literacy rate (0.69), and the percentage of people using the internet (0.86). while the highest loading value in 2nd factor for 1st cluster is the number of district receiving mobile library cars (0.7), the number of book collection (0.65), the number of libraries (0.93) and the number of digital library application (0.93). do it can be conclude that in 1st cluster has two factors, namely 1st factor represents the standard level of reading literacy, while 2nd factor represents reading literacy support facilities. In addition, can be concluded that the standard of reading literacy level in provinces that are member of 1st cluster can be increased if the IPLM, TGM, literacy rate, and the percentage of people using the internet are increased. Meanwhile, to improve the reading literacy of public in provinces that are members of 1st cluster, the mobile library car facilities, book collections, library buildings, and library digital applications must be improved in terms of quality and quantity.

In 2nd cluster for 1st factor, it can be seen that the highest loading values are the variables of the number of book collections (0.9), the number of libraries (0.92), TGM (0.8), the number of digital library applications (0.7), literacy rates (0.99), and the use of the internet (1). While the highest loading values in 2nd factor for 2nd cluster is IPLM (0.47) and the number of district receiving mobile library cars (0.98). so ut can be concluded that the factors that describe the reading literacy in provinces of 2nd cluster are 1st factor representing as reading literacy support facilities and 2nd factor representing as the standard level of reading literacy. Based on the result of factor analysis in each cluster, it can be concluded that efforts to increase the reading literacy in Indonesia are by increasing the standard of reading literacy level and supporting the facilities for reading literacy.

CONCLUSION

The k-means clustering creates two optimal clusters that describe the reading literacy index in Indonesia. 2nd cluster is some provinces that have the largest number of libraries, book collections, and district receiving mobile library cars in Indonesia. In addition, factor analysis in each cluster resulted two factors, namely the standard factor for reading literacy levels and supporting the facilities for reading literacy. So, it can be concluded that the way to increase reading literacy in two clusters of the area in Indonesia are by increasing the standard of reading literacy level and supporting the facilities for reading literacy. Based on this conclusion, hope the government can apply and be focus this treatment in each province in Indonesia to increase the literacy reading in Indonesia.

REFERENCES

Central Connecticut State University. The World's Most Literate Nations Ranked. Maret 2016. https://webcapp.ccsu.edu/?news=1767&data.

Hair, J. F. dan Anderson, R. E. (1998). Multivariate Data Analysis. New Jersey: Printice Hall.

Johnson, R. A. dan Wichern, D. W. (2002). Applied Multivariate Statistical Analysis. New Jersey: Prentice Hall.

Miller, John W and Mc.Kenna,Michael C. (2016). World literacy : how countries rank and why it matters. New York : Routledge.

PISA. 2018. Insight and Interpretation. U.S. Department of Education. Institute of Education Sciences, National Center for Education Statistics. Available at <PISA 2018 Insights and Interpretations FINAL PDF.pdf (oecd.org)>

Pusat Penelitian Kebijakan Pendidikan dan Kebudayaan. 2019. Indeks Aktivitas Literasi membaca 34 Provinsi. Jakarta : Puslitjakdikbud.

Satu Data Perpusnas. (2021, December). Retrieved from https://satudata.perpusnas.go.id/index.php/dashboard/.

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

Sutarno, NS. (2006). Perpustakaan dan Masyarakat. Jakarta: CV Sagung Seto.