Recognizing poverty pattern in Central Java using Biclustering Analysis

C A Putri*, R Irfani, and B Sartono
Department of Statistics, IPB University, Bogor, 16680, Indonesia

*E-mail: christiana_putri@apps.ipb.ac.id

Abstract. Poverty is a complex and multidimensional problem and becoming a development priority. In analyzing the pattern of poverty in a region, one of the statistical procedures that is usually used is the cluster analysis. However, it does not consider the different levels of performance by region in different characteristics at a particular time. In this study, an alternative approach, namely Cheng and Church’s biclustering algorithm, was used to simultaneously identify the poverty pattern in Central Java by region and poverty dimension variables. Using this algorithm, we found two biclusters with different characteristics. The first bicluster represents the general condition of poverty in Central Java, but they are better in labor and housing quality indicators. While the second bicluster is poorer in some indicators of labor, education, and housing quality than Central Java.

1. Introduction
Poverty is a complex and multidimensional problem and becoming a development priority. During the last two decades, the issue of population mobility from rural areas to cities or the so-called urbanization with poor status has become a discussion anywhere [1]. The Indonesian Government has great attention to the creation of a fair and prosperous society as contained in the fourth paragraph of the Constitution 1945. Development programs conducted during this time also always give great attention to the efforts of poverty alleviation because basically the development is done aims to improve the welfare of the community. One indicator of successful development is by increasing economic growth, with high economic growth expected to reduce poverty in an area. Since 2002, poverty levels in Indonesia have decreased, both in terms of numbers and percentages, except in 2006, September 2013 and March 2015. Poverty cannot be understood using just one dimension or one indicator. Simply, poverty is a lack of ability to achieve a minimum standard of living [2]. Housing, gasoline, electricity, education and toiletries are non-food commodities that have a considerable influence on the value of poverty line value both in urban and rural areas [3]. More than 50 percent of the poor population in Indonesia is on the island of Java, which is about 14 million people [4]. With several provinces has human development index (HDI) above the HDI Indonesia, which is 71.92 [5].

Central Java is an area with a fairly low human development index among other provinces on the island of Java, with a fairly high poverty rate. One indicator of poverty based on the multidimensional poverty index that education influences poverty. Among all the provinces in Java Island, Central Java is an area with the lowest mean years school, which is 8.03 [6]. In terms of life expectancy, Central Java has the second highest number after DI Yogyakarta which is 74.37 [7]. This shows that life expectancy in Central Java was not in line with mean years school. The result of the inconsistency between life
expectancy and mean years school caused the poverty rate in Central Java to be higher compared to other provinces on Java around 10.80 percent, which is the second highest after DI Yogyakarta which reached 11.70 [8]. On the other hand, Central Java has the second highest number of regencies after East Java, amounting to 35 districts. The percentage of poor people in Indonesia in March 2019 was around 9.41 percent or 25.14 million people. Decreased by 1.8 million people (9.82 percent) compared to March 2018. As one of the major provinces in Indonesia, Central Java has a Headcount Index of around 10.80 percent in March 2019, down from March 2018, which amounted to 11.32 percent.

Poverty can be shown as a multidimensional problem because it is related to economic, social, cultural, political inability and participation in society. Poverty has a broader meaning than just lowering one's income or consumption level from measured welfare standards such as minimum calorie requirements or poverty lines, but poverty has a deeper meaning because it is related to the inability to reach non-income aspects such as access to minimum needs; health, education, clean water and sanitation. The complexity of poverty is not only related to understanding and dimensions but also related to the method used to measure the poverty line. The Multidimensional Poverty Index (MPI) was used since 2010 which is a multidimensional poverty measure. MPI shows a deeper poverty structure not only in income or consumption but also in multidimensional. MPI measures poverty from various dimensions such as the health dimension, the education dimension, the labor dimension, and the housing quality dimension. The MPI dimension is often used in research on poverty [9] [10]. With MPI, the local government will describe deeper poverty and its relation to the basic needs approach that has been used by BPS.

One of the statistical procedures that are being used to determine districts/cities grouping according to their poverty characteristics is known as the cluster analysis. Cluster analysis may be used to identify groups of subject that are homogeneous to each other based on similarity measurements. In most poverty studies, a common clustering analytical technique that is often used is the one way clustering method. One of the weaknesses of one-way clustering is that it only considers region’s values that are similar to each other. However, in reality some regions may only perform poorly in some and not all dimension of poverty. We are interested to identify these weak regions so that extra attention can be provided to them in order to enhance their achievement in the future. A point of concern regarding the use of one-way clustering is that there is a limitation in the way it detects data of performance which is only one sided. Therefore, in overcoming this problem, we use the biclustering method (Cheng and Church algorithm) in determining the regions performance specifically in certain dimension of poverty. The use of biclustering method permits us to link corresponding data in clustering simultaneously between districts/cities and poverty characteristics. Therefore this study aims to identify the poverty pattern in Central Java by region and poverty dimension variables simultaneously using biclustering analysis.

2. Material and methods

2.1. Data and variables
We used data from National Socio-Economic Survey (Susenas) by the Central Beurau of Statistics Indonesia in 2019. Susenas provides data for socio-economic aspects and fulfillment of life necessities such as clothing, food, housing, health, safety and employment opportunities. The observation unit is all districts/cities in Central Java. The poverty variables used in this research are taken from multidimensional poverty indicator, more details are shown in Table 1.
Table 1. Multidimensional poverty indicator

| Dimension             | Variable                                                                 |
|-----------------------|--------------------------------------------------------------------------|
| **Poverty**           | Headcount index (\(P_0\))                                               |
|                       | Poverty gap index (\(P_1\))                                             |
|                       | Poverty severity index (\(P_2\))                                        |
| **Health**            | Life expectancy (\(X_1\))                                               |
|                       | Percentage of poor households using clean water sources for drinking (\(X_2\)) |
|                       | Percentage of poor households with latrines owned (\(X_3\))             |
| **Labor**             | Percentage of poor people aged 15 years and over who do not work (\(X_4\)) |
|                       | Percentage of poor people age 15 and older who work in the agricultural sector (\(X_5\)) |
|                       | Percentage of poor people age 15 and older who work in the informal sector (\(X_6\)) |
| **Education**         | Mean years school of people age 15 and older (\(X_7\))                 |
|                       | Percentage of poor people age 15 and older who do not complete primary school (\(X_8\)) |
|                       | Literacy rate of poor people aged 15-55 years old (\(X_9\))             |
| **Housing quality**   | Percentage of households with bamboo/other as wall main material (\(X_{10}\)) |
|                       | Percentage of households with earth/other as floor main material (\(X_{11}\)) |
|                       | Percentage of households that use kerosene, charcoal, wood / others as the main cooking fuel (\(X_{12}\)) |

2.2. Methods
Two way data assumes that certain objects only have characteristics in certain rows or columns. Identification of certain related row subgroups and column subgroups is the main objectives of two way clustering. This study discusses a two way clustering approach that connects row subgroups with column subgroups simultaneously on the data matrix. The main idea of this approach is to partition a matrix into a submatrix with a singular value decomposition approach, then the grouping is carried out simultaneously on the submatrix so that a group of rows and a group of columns is related to each other or is called as bicluster analysis.

By definition, biclustering refers to the clustering algorithm that performs a simultaneous row and column clustering. This method is an unsupervised technique data mining which groups data in a two dimensional matrix that performs a simultaneous row and column clustering. Biclustering has become a popular approach to analyzing high-dimensional biological data sets and multiple algorithms [11]. Compared to the one way clustering, bicluster analysis has robustness and more informative.

In cluster analysis, matrix \(A\) with size \(n \times m\) can be written as follows:

\[
\begin{array}{cccc}
  y_1 & \cdots & y_i & \cdots & y_m \\
  x_1 & a_{11} & \cdots & a_{1i} & \cdots & a_{1m} \\
  \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
  x_j & a_{ij} & \cdots & a_{ij} & \cdots & a_{mj} \\
  \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
  x_n & a_{1n} & \cdots & a_{in} & \cdots & a_{mn} \\
\end{array}
\]
with X as objects, Y as variables and entries $a_{ij}$. Bicluster analysis tries to find subgroup $A_{IJ}$ of objects $I = \{i_1, \ldots, i_k\}, k \leq n, I \subset X$ which are as similar as possible to each other on a subset of variables $J = \{j_1, \ldots, j_l\}, l \leq m, J \subset Y$ and as different as possible to the remaining objects and variables.

The concept of a bicluster is a subset of rows and a subset of columns with a high similarity score [12]. The similarity score between the object in bicluster is measured by mean squared residue (MSR). MSR measures the deviation of value in a bicluster from the mean value of the bicluster [13]. Cheng and Church’s (CC) algorithm tries to find a maximum bicluster with MSR lower than a threshold.

Defined the data as a matrix $A(X,Y)$, $X$ as the rows set and $Y$ as the columns set. A bicluster is represented as a sub-matrix $(I,J)$, with $I \subset X$ and $J \subset Y$, and has MSR below a certain threshold ($\delta$), where $\delta \geq 0$. It is called a $\delta$-bicluster. The residue score of element $a_{ij}$ in the bicluster is

$$a_{ij} - a_{ij} - a_{lj} + a_{ij}$$

and the MSR score of a bicluster is:

$$MSR(I,J) = \frac{1}{|I||J|} \sum_{i \in I, j \in J} (a_{ij} - a_{ij} - a_{ij} + a_{ij})^2$$

where $a_{ij}$ is the mean of the $i$th row, $a_{ij}$ is the mean of the $j$th column, and $a_{ij}$ is the mean of all elements in the bicluster

$$a_{ij} = \frac{1}{|I|} \sum_{j \in J} a_{ij}, \quad a_{ij} = \frac{1}{|J|} \sum_{i \in I} a_{ij} \quad a_{ij} = \frac{1}{|I||J|} \sum_{i \in I, j \in J} a_{ij}$$

In the CC algorithm, the quality of bicluster is depend on the residue and the volume of the bicluster. The lower the residue and/or the larger the volume of the bicluster, the better its quality. Therefore residual to volume mean ratios were used to compare the qualities of the different bicluster [14]:

$$\frac{1}{n} \sum_{i=1}^{n} \frac{MSR_i}{Volume_i}$$

2.3. Tools
This research use package biclust for biclustering analysis and package sp for the spatial map, which executed using Rstudio version 1.2.1335.

3. Result and discussion
Figure 1 show the heatmap of poverty structure in Central Java. It visualizes regional and variable group based on their correlation values with dendograms. Districts/cities or variables that are correlated will be in one group. From the heatmap in Figure 1 it can be seen that there are 2 large groups namely life expectancy ($X_1$), percentage of poor households using clean water sources for drinking ($X_2$), percentage of poor households with latrines owned ($X_3$), and the literacy rate of poor people aged 15-55 years old ($X_9$) as the first group; and the remaining variables as the second group. The regional dendrogram in Figure 1 also consists of two large groups, the first regional group consist of 12 districts (upper cluster) and the remaining districts/cities as second regional group (bottom cluster). For dark cells, it shows areas with high variable values. The area in the upper cluster tends to be darker than the area in the bottom cluster, this shows that the variable value in the upper cluster region is higher than the area in the bottom cluster. Except for the percentage of poor people aged 15 years and over who do not work ($X_4$), the values in the upper cluster are lower than the bottom cluster.
In the bicluster approach, we run some parameters and do manual tuning to choose the best parameter based on the average ratio of residue over volume (MSR/V). CC algorithm tries to get bicluster as large as possible with a low MSR value; therefore, the best parameter should be the one that generates the smallest MSR/V. From the results of manual tuning in Table 2, it can be seen that although $\delta$ is a threshold parameter, the $\delta$ value is not linear to the MSR/V. There is a positive relationship between $\delta$ value and MSR/V only in $\delta$ value 0.1 - 0.3 and 0.4 - 0.6, where greater $\delta$ value generates greater MSR/V score. The $\delta$ values 0.7 – 1 produces the smallest MSR/V value but only produces one bicluster. Based on MSR/V score and the number of bicluster generated, $\delta = 0.4$ is chosen as the best parameter. In interval 0.4 – 0.6, the number of bicluster is convergent at 2 bicluster and the smallest MSR/V score is produced when $\delta = 0.4$.

Table 2. The result of parameter tuning

| $\delta$ | Number of bicluster | MSR/V   |
|--------|---------------------|---------|
| 0.1    | 5                   | 0.002328|
| 0.2    | 4                   | 0.003288|
| 0.3    | 3                   | 0.004461|
| **0.4**| **2**               | **0.002079**|
| 0.5    | 2                   | 0.002860|
| 0.6    | 2                   | 0.007687|
| 0.7    | 1                   | 0.001262|
| 0.8    | 1                   | 0.001262|
| 0.9    | 1                   | 0.001262|
| 1      | 1                   | 0.001262|
The detail result of CC algorithm with $\delta = 0.4$ is shown in Table 3. Using $\delta = 0.4$ as parameter, we find 2 bicluster which is quite similar to the heatmap result in Fig.1. Each of district/city in Central Java is covered exactly into one bicluster, except the district of Tegal that is not grouped in any bicluster. Most districts/cities are included in first bicluster 1. First bicluster consist of 24 districts/cities which they had same characteristic in 14 variables, i.e variable $X_1$ and $X_2$ related to health dimension, variable $X_4$, $X_5$, and $X_6$ related to labor dimension, variable $X_7$, $X_8$, and $X_9$ related to education dimension, and variable $X_{10}$, $X_{11}$, and $X_{12}$ related to housing quality dimension. Then the second bicluster has 10 districts which they had same characteristics in 13 variables, i.e variables $X_1$ and $X_3$ related to health dimension, variable $X_4$, $X_5$, and $X_6$ related to labor dimension, variable $X_7$, $X_8$, and $X_9$ related to education dimension, and variables $X_{10}$ and $X_{12}$ related to housing quality dimension. Meanwhile variable $P_0$, $P_1$, and $P_2$ is included in both bicluster. It means that the pattern of absolute poverty measurements is similar with the pattern of poverty determinants in each bicluster. This result is in line that the theoretically supported dimensions of poverty are all related, and able to reveal more realistically the overall poverty status of households [15]. Different from previous research, that the MPI cannot describe the pattern of poverty in East Java [10]. Figure 2 show the bicluster map of poverty in Central Java. It can be seen that in general close districts/cities tend to be grouped into the same bicluster, or they have similar pattern of poverty. This implied that spatial effect is also influence the poverty condition in a district/city.

| Table 3. Details of each bicluster |
|-----------------------------------|
| **Bicluster 1**                  | **Bicluster 2**                  |
| Variables: | Variables: |
| $X_1$ $X_2$ $X_4$ $X_5$ $X_6$ $X_7$ $X_8$ $X_9$ $X_{10}$ $X_{11}$ $X_{12}$ $P_0$ $P_1$ $P_2$ | $X_1$ $X_3$ $X_4$ $X_5$ $X_6$ $X_7$ $X_8$ $X_9$ $X_{10}$ $X_{12}$ $P_0$ $P_1$ $P_2$ |
| Kab Cilacap 9 | Kab Sragen 17 | Kab Pemalang 1 | Kab Purworejo 1 |
| Kab Banyumas 10 | Kab Rembang 18 | Kab Brebes 2 | Kab Wonosobo 2 |
| Kab Purbalingga 11 | Kab Pati 19 | Kota Magelang 3 | Kab Magelang 3 |
| Kab Banjarnegara 12 | Kab Kudus 20 | Kota Surakarta 4 | Kab Boyolali 4 |
| Kab Kebumen 13 | Kab Jepara 21 | Kota Salatiga 5 | Kab Wonogiri 5 |
| Kab Klaten 14 | Kab Demak 22 | Kota Semarang 6 | Kab Grobogan 6 |
| Kab Sukoharjo 15 | Kab Kendal 23 | Kota Pekalongan 7 | Kab Blora 7 |
| Kab Karanganyar 16 | Kab Pekalongan 24 | Kota Tegal 8 | Kab Semarang 8 |
|                     |                   |                  | Kab Temanggung 9 |
|                     |                   |                  | Kab Batang 10 |

Figure 2. Biclusters map of poverty in Central Java
Table 4 shows the result of two samples t-test for comparing two means to show the characteristics of each bicluster. The mean of each bicluster was compared with Central Java mean. The test result show that bicluster 1 and bicluster 2 are different in some characteristics. Bicluster 1 has lower mean of variable X₅ than Central Java, while bicluster 2 has higher mean than Central Java. In general, bicluster 1 has no difference with Central Java except variable X₅, X₆, and X₁₂ which is significantly lower than Central Java. Compared to Central Java, districts and cities in bicluster 1 have lower percentage of poor people age 15 and older who work in agricultural sector (X₅) and informal sector (X₆), and the percentage of households that use kerosene, charcoal, wood/others as the main cooking fuel (X₁₂). From Table 4, we can see that there is no significant difference in absolute poverty measurements (P₀, P₁, and P₂) between bicluster 1 and province rate. It means that bicluster 1 represents the general condition of poverty in Central Java but they are better in some indicators of labour and housing quality.

| Variable | Mean | Stdev | Variable | Mean | Stdev | Variable | Mean | Stdev |
|----------|------|-------|----------|------|-------|----------|------|-------|
| X₁       | 74.987 | 1.997 | X₁       | 74.612 | 1.322 | X₁       | 74.777 | 1.88 |
| X₂       | 74.343 | 13.252 | X₃       | 86.01 | 8.827 | X₂       | 73.161 | 13.522 |
| X₄       | 42.481 | 4.569 | X₄       | 33.174 | 4.072 | X₃       | 85.75 | 8.524 |
| X₅       | 31.624 | 3.39 | X₅       | 38.287 | 6.066 | X₄       | 40.27 | 6.797 |
| X₆       | 31.624 | 8.601 | X₆       | 52.089 | 3.291 | X₅       | 21.838 | 14.115 |
| X₇       | 8.013 | 1.339 | X₇       | 7.227 | 0.543 | X₆       | 37.036 | 12.146 |
| X₈       | 25.777 | 9.453 | X₈       | 32.504 | 7.142 | X₇       | 7.755 | 1.201 |
| X₉       | 98.212 | 1.864 | X₉       | 97.275 | 1.864 | X₈       | 27.958 | 9.335 |
| X₁₀      | 0.818 | 0.771 | X₁₀      | 1.073 | 0.976 | X₉       | 97.796 | 2.023 |
| X₁₁      | 7.814 | 6.418 | X₁₂      | 27.009 | 8.842 | X₁₀      | 0.888 | 0.818 |
| X₁₂      | 10.58 | 3.253 | P₀       | 10.643 | 2.568 | X₁₁      | 10.275 | 9.908 |
| P₀       | 10.457 | 3.767 | P₁       | 1.042 | 0.572 | X₁₂      | 15.106 | 10.796 |
| P₁       | 1.563 | 0.67 | P₂       | 0.166 | 0.131 | P₀       | 10.429 | 3.404 |
| P₂       | 0.345 | 0.178 |          |      |      | P₁       | 1.388 | 0.681 |
|          |      |      |          |      |      | P₂       | 0.286 | 0.184 |

*a significant at α = 0.05 compare to Central Java (mean)

Bicluster 2 consist of fewer districts compared to bicluster 1. It has only 10 districts with 13 similarity variables and 6 variables are significantly different with Central Java. From Table 4, it can be seen that the member of bicluster 2 have higher percentage of poor people age 15 and older who work in agricultural (X₅) and informal sector (X₆) than Central Java but they have lower percentage of unemployment (X₄). In the other hand, the mean years school (X₇) of districts in bicluster 2 is significantly lower and the percentage of households that use kerosene, charcoal, wood/others as the main cooking fuel (X₁₂) is higher than Central Java. It means that districts in bicluster 2 is poorer in mean years school as indicator of education dimension and percentage of households that use kerosene, charcoal, wood/others as the main cooking fuel as indicator of housing quality. The absolute poverty measurements in bicluster 2 is similar with Central Java except the poverty severity index (P₂) which is better than province rate.

4. Conclusions

The application of the CC biclustering algorithm to the poverty structure in Central Java produced 2 bicluster. Bicluster 1 consist of 24 districts/cities with 14 variables and bicluster 2 consists of 10 districts with 13 variables. Variable P₀, P₁, and P₂ as absolute poverty measurements is included in both bicluster.
It means that the pattern of absolute poverty measurements is similar with the pattern of poverty determinants in each bicluster. The pattern of poverty dimension in Central Java that forms bicluster involves only several indicator variables include in health, labor, education, and housing quality dimension. From the pattern found in each bicluster, the government is expected to develop poverty alleviation policies at various dimensions appropriately.

References
[1] Lucci P, Bhatkal T and Khan A 2018 Are we underestimating urban poverty?. World development, (103) 297-310.
[2] Liu Y, Liu J and Zhou Y 2017 Spatio-temporal patterns of rural poverty in China and targeted poverty alleviation strategies. Journal of Rural Studies, (52) 66-75.
[3] [BPS] Badan Pusat Statistik 2019 Berita Resmi Statistik: Profil Kemiskinan Indonesia Maret 2019. Jakarta (ID): Badan Pusat Statistik.
[4] [BPS] Badan Pusat Statistik 2019 Data dan Informasi Kemiskinan Kabupaten/Kota Tahun 2019. Jakarta (ID): Badan Pusat Statistik.
[5] [BPS] Badan Pusat Statistik 2020 [Metode Baru] Indeks Pembangunan Manusia menurut Provinsi. [internet]. Available at: https://www.bps.go.id/indicator/26/494/1/metode-baru- indeks-pembangunan-manusia-menurut-provinsi.html
[6] [BPS] Badan Pusat Statistik 2020 Rata-rata Lam Sekolah Penduduk Umur ≥ 15 Tahun Menurut Provinsi, 2018-2020. [Internet]. Available at: https://www.bps.go.id/indicator/28/1429/1/rata-rata-lama-sekolah-penduduk-umur-15-tahun-menurut-provinsi.html
[7] [BPS] Badan Pusat Statistik 2020 [Metode Baru] Umur Harapan Saat Lahir (UHH) (Tahun), 2019-2020. [internet]. Available at: https://www.bps.go.id/indicator/26/494/1/metode-baru-indeks-pembangunan-manusia-menurut-provinsi.html
[8] [BPS] Badan Pusat Statistik 2020 Persentase Penduduk Miskin Menurut Provinsi (Persen), 2019-2020. [internet]. Available at: https://www.bps.go.id/indicator/23/621/1/persentase-penduduk-miskin-menurut-kabupaten-kota.html
[9] Anuraga G and B W Otok 2013 Pemodelan Kemiskinan di Jawa Timur dengan Structural Equation Model – Partial Least Square (Poverty Modelling in East Java Using Structural Equation Model-Partial Least Square). Statistika, 1(2): 121-130.
[10] Yuniarto B and Kurniawan R 2017 Understanding structure of poverty dimensions in east java: Bicluster approach. Signifikan: Jurnal Ilmu Ekonomi 2017; 6(2): 289–300.
[11] Yanjie Z and Hongbo S 2018 Application of biclustering algorithm to extract rules from labeled data. International Journal of Crowd Science.
[12] Cheng Y and Church GM. 2000. Biclustering of expression data. Ismb; vol. 8: 93–103.
[13] Yang J, Wang W and Wang H, Yu P 2002. δ -Clusters: Capturing Subspace Correlation in a Large Data Set. Proc. 18th IEEE Int’l Conf. Data Eng. p.517-528.
[14] Chakraborty A and Maka H. 2005. Biclustering of gene expression data using genetic algorithm. 2005 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology. p.1–8.
[15] Wagle U 2005 Multidimensional poverty measurement with economic well-being, capability, and social inclusion: a case from Kathmandu, Nepal. Journal of Human Development, 6(3), pp.301-328