Countries population determination to test rice crisis indicator at national level using $k$-means cluster analysis

Y. Hidayat*, T. Purwandari2, Sukono3, Y.D. Ariska4

1,2Department of Statistics, Universitas Padjadjaran, Indonesia
3Department of Mathematics, Universitas Padjadjaran, Indonesia
4Graduate School of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran, Indonesia

*E-mail: yuyun.hidayat@unpad.ac.id

Abstract. This study aimed to obtain information on the population of the countries which is have similarities with Indonesia based on three characteristics, that is the democratic atmosphere, rice consumption and purchasing power of rice. It is useful as a reference material for research which tested the strength and predictability of the rice crisis indicators Unprecedented Restlessness (UR). The similarities countries with Indonesia were conducted using multivariate analysis that is non-hierarchical cluster analysis $k$-Means with 38 countries as the data population. This analysis is done repeatedly until the obtainment number of clusters which is capable to show the differentiator power of the three characteristics and describe the high similarity within clusters. Based on the results, it turns out with 6 clusters can describe the differentiator power of characteristics of formed clusters. However, to answer the purpose of the study, only one cluster which will be taken accordance with the criteria of success for the population of countries that have similarities with Indonesia that cluster contain Indonesia therein, there are countries which is sustain crisis and non-crisis of rice in 2008, and cluster which is have the largest member among them. This criterion is met by cluster 2, which consists of 22 countries, namely Indonesia, Brazil, Costa Rica, Djibouti, Dominican Republic, Ecuador, Fiji, Guinea-Bissau, Haiti, India, Jamaica, Japan, Korea South, Madagascar, Malaysia, Mali, Nicaragua, Panama, Peru, Senegal, Sierra Leone and Suriname.

1. Introduction
In 2007-2008 recorded as the world rice crisis [5]. It is marked with an extremely increasing of rice prices in the world market, it was reached 2.852 times higher between October 2007 and May 2008 [4]. Not a few countries in Asia and Africa affected by the rice crisis. Indonesian Government should be responsive to the situation that afflicted these countries. Why? Because the rice crisis can also afflict Indonesia. It is because majority of Indonesian consume the rice as one of their staple food.

Indonesian Government needs a predictor to look for the potential of rice crisis in Indonesia in the future. Therefore, some researchers do studies in order to know the strength and predictability of indicator called Unprecedented Restlessness (UR) to detect the occurrence of rice crisis in the future. However, there are still less of data to describe the economic situation In Indonesia to predict the rice crisis in the future, it creates fears of uncertainty to look for the dynamics of rice crisis in Indonesia. Therefore, the researchers used the data of countries that assumed have similar characteristic to
Indonesian’s without do the testing first. The effect is not known prediction accuracy of the result of the studies about the occurrence of rice crisis in Indonesia [8].

Therefore, it needs a testing to form a group that consists the countries which have similarities to Indonesian’s. In this study, the similarity of countries in Indonesian is measured by 3 characteristics (democratic atmosphere, rice consumption, and restlessness or rice purchasing ability) in each countries [7].

2. Literature Reviews

In this study, there are 3 based characteristics that formed a group or population of countries which is have similarities to Indonesian’s. The importance of the measured similarities is basic formation to form the population of countries which is have similarities to Indonesian’s based on 3 characteristics, so the formed population can be used as an object of study to test the indicator of occurrence of rice crisis in the future.

2.1. The Similarity of Characteristics

2.1.1 Democratic Atmosphere

The advanced study Polity IV Project always develop coding to know the characteristics of authority in countries (world’s systems) to compare and do a quantitative analyse which is illustrated by score called Polity. Polity score that produced by this project has been proved by the researchers years by years. This score is the democratic measurement of each country. The scale of this score ranges between -10 and +10, which is -10 shows the very authoritative government’s system and +10 shows the very democratized government’s system [10].

2.1.2 Percent Rice Consumption

Rice is a staple food for almost half of population in the world. Besides, more than 90% of population in Asia consume rice as their staple or main food.

Percent Rice Consumption is a proportion of rice consumption to total consumption of staple food in a country [7].

\[
\%C = \frac{RC}{TC} \times 100\%
\]

Note:
\(R=\text{Restlessness}; P=\text{Rice Prices (US$/kg)}; C=\text{Rice Consumption (kg/capita)}; I=\text{Income per capita (US$/capita)}.

2.1.3 Restlessness

This characteristic measured with rice purchasing ability per capita in a country, this is the formula below [7]:

\[
R = \frac{P}{I} C
\]

Note:
\(R=\text{Restlessness}; P=\text{Rice Prices (US$/kg)}; C=\text{Rice Consumption (kg/capita)}; I=\text{Income per capita (US$/capita)}.

2.2. Multivariate Analysis

Multivariate analysis is an application of methods that can be used when some measurement (1 and more) obtain from each individuals or observations. This measurement called variables. In multivariate analysis, all variables measured simultaneously in each sample unit [9]. Basically, in multivariate analysis there are 2 classifications of relationship between variables, that is dependency relationship and interdependency relationship. Dependencies technique occurs when there is a
dependency relationship between variables, and Interdependencies occurs when there are no different between dependent variable and independent variable [2].

In this study, interdependencies technique was used, because there is no different of dependent or independent variables. Therefore, there are some possibilities of application of multivariate analysis that used to form a group from observations, such as multidimensional scaling analysis, correspondence analysis, biplot analysis, and cluster analysis [9].

2.2.1 Dependencies Technique

Dependencies technique of multivariate analysis is a method to explain or predict one or more criteria measurement based on a set of predictor variables (independent variables). Multivariate analysis with dependencies technique can be used when there are independent and dependent variables. Some methods that classified to dependencies technique are multiple regressions, discriminant analysis, canonic analysis, and others [2].

2.2.2 Interdependencies Technique

Interdependencies technique of multivariate analysis is a method to give a more evident description about data structure by simplify the complex data structure (data reduction). This technique is used when there is no different between dependent variables and independent variables (predictors). Multivariate method that consisted in this method are multidimensional scaling analysis, correspondence analysis, biplot analysis, and cluster analysis.

Multidimensional scaling analysis is a multivariate method that used to determine the position of an observation to another observation based on the similarities value. The advantage of this method, can be used for metric and non-metric data structure while the weakness of this method is plot data structure of observations became not evident when there is a data point with big dimension [9]. The advantage of correspondence analysis method is it flexible enough used for a big dimension of data structure, but the weakness of this method is it only can be used for non-metric data structure (nominal and ordinal) [9]. The advantage of biplot analysis, it is capable to show directly the dominant variables of group of observations formed [13]. However, the weakness of this method, when do a big dimension reduction to 2 dimensions, then it can be effect to decreasing of information in biplot.

2.3. Cluster Analysis

Cluster analysis is a troubleshooting methods based on multivariate data. This method used to determine the cluster to group from some observations that have high similarities inside the cluster, but they have high dissimilarities between the clusters. This method is different with another grouping technique like discriminant analysis, which in cluster analysis, it shows unknown groups so that the objective of this method is to form group (clusters). To form clusters using this method, the similarities technique is the first thing that mostly seen, that is based on the closeness among the pair of each observations. The other methods can also be used to determine the centroid of some clusters [14].

There are 2 methods of grouping in cluster analysis, hierarchical and non-hierarchical technique (partition). Hierarchical methods begin with grouping observations by the nearest distance among all pair of observations, it begin with the big number of clusters then become smaller as a clusters merge to another clusters. It is different from non-hierarchical (partition) method, that used by determine the number of clusters first, then it finds the observations that have the nearest distance to the centroid of clusters [14].

The closeness or distance that usually used in cluster analysis between 2 observations in p dimensions (variables) is Euclidean Distance [14].

\[ d(x, y) = \sqrt{(x - y)'(x - y)} = \sqrt{\sum_{j=1}^{p} (x_j - y_j)^2} \]  

(3)

where \( x = (x_1, x_2, \ldots, x_p)' \) and \( y = (y_1, y_2, \ldots, y_p)' \).
Hierarchical Method
Cluster analysis with this method can be done with 2 approaches, agglomerative and divisive approaches. Agglomerative approach begin with each observations as clusters then the nearest observations will be merged first and forward merge into another clusters that have the nearest distance each other, and then until the number of clusters become smaller and the member of each clusters become larger.

Whereas, in the divisive approach, it begin with dividing the population into 2 sub-groups so the observations consisted in each cluster have the furthest distance each other, then they will divided again into the next sub-groups, continued until it form the largest number of clusters but the smallest number of member of each clusters. The final result of this method formed a dendogram [9].

Non-Hierarchical Method (k-Means)
Cluster analysis with this method can be done with determine the number of clusters to be formed first. This method begins with one of two approaches that is initial partition of observations into groups and first determining of centroids of each clusters. k-Means method is done by grouping items or observations into cluster that have the nearest centroid [9].

There are some steps below that can be done in this cluster analysis with non-hierarchical methods:

- Partitioning the observations randomly into k initial clusters.
  The first thing that should be done is determining the number of initial clusters, that is k clusters, so we get the centroid of each clusters [14].
  The centroid value of each clusters calculated by below formula [15]:

\[ \mu_k = \frac{1}{N_k} \sum_{q=1}^{N_k} x_q \]  

(4)

Note:
\[ \mu_k \] : Centroid of cluster-k
\[ N_k \] : The number of members of cluster-k
\[ x_q \] : Observations-q in cluster-k

- Groups the observations into clusters that have the nearest centroid.
  Distance that used in this analysis is Euclidean distance, both standardized and non-standardized. Recalculate the centroid value of each cluster that had a member change (addition or subtractions of cluster members) [9].

- Performs the repetition steps 2 until it impossible to occur the change of cluster members [9].

The advantage of k-Means cluster compared to hierarchical method is it can overcome the large number of data. k-Means cluster is method that used in this study because the similarities among the observations with metric data structure and interdependencies technique (there is no different between dependent and independent variables).

2.4. Differentiator Power of Variables
To measure the ability of the three variables in differentiate the formed clusters simultaneously (multivariate) then statistic testing that can be used is T²Hotelling analysis for 2 populations (cluster) and MANOVA for advanced cases more than 2 populations (clusters).

2.4.1 T²Hotelling
In univariate case (one variable), the analysis to test the mean difference among 2 populations used is t-test analysis. The statistical testing that used as the formula below [14]:

\[ t = \frac{\bar{y}_1 - \bar{y}_2}{s_{gb} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \]  

(5)

From the formula above, it notes that \( \bar{y}_1 \) and \( \bar{y}_2 \) is the means value of variables of population 1 and 2. For case using some variables in 2 populations, t-test analysis can not be performed, so that it will
need multivariate analysis to test the difference of 2 mean vectors using $T^2$ Hotelling which is can perform an analysis to some variables of 2 populations simultaneously [14].

$$T^2 = (\bar{y}_1 - \bar{y}_2)' \left( \frac{1}{n_x} + \frac{1}{n_y} \right) S_{gab}^{-1} (\bar{y}_1 - \bar{y}_2)$$  \hspace{1cm} (6)

From the equation above, $\left( \frac{1}{n_x} + \frac{1}{n_y} \right) S_{gab}$ as covarians to $(\bar{y}_1 - \bar{y}_2)$.

2.4.2 MANOVA

Multivariate analysis of variance is an advanced method of univariate technique ANOVA to predict the difference between some groups (>2 populations). MANOVA is capable to performs the testing to determine the difference between >2 populations by using some variables simultaneously. If there are some multiple variate then this analysis will give the dimension of differences that will differentiate between some groups better.

To perform the testing for differences means vector between k populations note in a form null hypotheses $H_0: \mu_1 = \mu_2 = \cdots = \mu_k$ using statistics Wilks Lambda testing with statistics $F$ approach.

$$\Lambda = \frac{|E|}{|E|+|H|}$$  \hspace{1cm} (7)

H is sum squares matrix between groups and E is sum squares matrix in the groups.

| Source of Variation | Sum Squares Matrix | Degree of Freedom |
|---------------------|--------------------|------------------|
| Effect              | $H = \sum_{l=1}^{k} \frac{1}{n_l} (\bar{y}_l - \bar{y}) (\bar{y}_l - \bar{y})'$ | k-1 |
| Residual            | $E = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\bar{y}_{ij} - \bar{y}_l)(\bar{y}_{ij} - \bar{y}_l)'$ | $\sum_{i=1}^{k} n_i - k$ |
| Total               | $H + E = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\bar{y}_{ij} - \bar{y})(\bar{y}_{ij} - \bar{y})'$ | $\sum_{i=1}^{k} n_i - 1$ |

2.4.3 Evaluation of Grouping Result

After performing the grouping process with cluster analysis, then we need to evaluate the result of the grouping of one cluster. This occurs because there is only one cluster that will be taken as the answer of the problem in this study that is the population of countries which is have similarities to Indonesian’s. The objective of this evaluation is that population to be right on target as the objectives of this study. There are things that should be noted as criteria of success of this grouping, the population of countries which is has similarities to Indonesian’s [7]:

- There is Indonesia inside of that group. The existence of Indonesia inside of the cluster or group is very important as a criteria of success, because we will look for the countries which is have the similarities to Indonesia’s so Indonesia should be there in the cluster.
- Inside of the cluster or group, there are both countries which is experience the rice crisis in 2008 and countries which did not experience the rice crisis in 2008.
- The cluster or group is the group that have the most number of cluster’s members. The more countries inside the cluster or group which is have the similarities to Indonesian’s the better to perform the study which is aim for testing of indicator called Unprecedented Restlessness (UR) as indicator of rice crisis.
3. Methodology

3.1. Data and Variables
The data sources for this study is a secondary data of 38 countries which is obtained by taking the countries that consumes rice as their staple or main food, and also rice cannot be substitute with another food [7]. Variables in this study are below:
X1: Democratic Atmosphere
X2: Percent Rice Consumption
X3: Restlessness

| No | Countries | Polity | Rice Consumption (%) | Restlessness |
|----|-----------|--------|-----------------------|--------------|
| 1  | Bangladesh| 6      | 76.9031               | 0.0399934482524908 |
| 2  | Brazil    | 8      | 21.7572               | 0.00173088466516339 |
| 3  | Cambodia  | 2      | 79.6210               | 0.0572555778572883 |
| ...| ...       | ...    | ...                  | ...          |
| 38 | Vietnam   | -7     | 82.0934               | 0.0361490469875234 |

3.2. The Steps of Study

3.2.1 Standardized Variables
The data standardizing is performed to overcome the variables that have the difference measurement scaling. In this study, there are 3 variables which is have the difference of unit so it it needs to perform the standardizing of variables that is using normal standard (Z score).

The most common form that used to perform data standardizing is to convert each variables to standard score (Z score) by subtracting the mean of variables from the observation value then divide it to standard deviation of each variables. The most common way of this method is function normalized distance [2]:

Note:
\( Z = \text{Normalize Standard Value} \); \( X = \text{the valueof observations} \); \( \mu = \text{Means of population} \); \( \sigma = \text{standard deviation of population} \)

3.2.2 Cluster Analysis: k-Means Method
There are some steps that will be performed:
- Partitioning each observations (countries) to \( k \) clusters randomly
- Calculate the centroid value of each clusters
- Move the observations (countries) to the nearest cluster
- Perform the step c repeatedly until there is no possibility of change of cluster’s members.

3.2.3 Differentiator Power of Variables
To look for the differentiator power of 3 variables, then it uses the test of T²Hotelling (2 populations) and MANOVA (>2 populations).
- The formulation of hypotheses
  \( H_0 \quad : \mu_1 = \cdots = \mu_k \)
  There is no difference of means of k- populations
  (3 variables are not capable to differentiate the formed populations)
  \( H_1 \quad : \text{There is difference of means of k- populations} \)
(3 variables are not capable to differentiate the formed populations)

- Determine the $\alpha$ value
- Statistics Testing
  - $T^2$ Hotelling and Wilks Lambda (Statistics $F$ approach)
- Testing Criteria
  - $T^2$ Hotelling : Reject $H_0$ if $F > F_{\alpha,p,n_1+n_2-2}$, acceptance for others.
  - MANOVA : Reject $H_0$ if $F > F_{\alpha,p,n_1+n_2-2}$, acceptance for others.

3.2.4 Evaluation of Grouping Result

As described in Chapter 2, there are 3 criteria of success of population groupings which is have similar characteristics with Indonesian’s, it is becoming a benchmark for the success of getting the appropriate population and meet the goals as the solution for the problem in this study.

4. Result

4.1. Standardized Variables

| No | Countries  | Polity | % Rice Consumption (%) | Restlessness |
|----|------------|--------|------------------------|--------------|
| 1  | Bangladesh | 0.44493| 0.031265               | 1.975335     |
| 2  | Brazil     | 0.7475 | -0.60989               | -1.1743      |
| 3  | Cambodia   | -0.1643| 0.320521               | 2.130567     |
| ...| ...        | ...    | ...                    | ...          |
| 38 | Vietnam    | -1.5352| -0.03315               | 2.271773     |

From the result of standardized scores as the Table 4.1, then the next is perform the grouping for countries using the K-Means clustering analysis.

4.2. k-Means Clustering Analysis & Differentiator Power of Variables

In choosing the number of clusters that will be performed, then it begin by form the smallest number of cluster that is 2 clusters.

| Table 4 Cluster's Members 2 Klaster | Table 5 Differentiator Power 2 Klaster |
|-------------------------------------|----------------------------------------|
| 2 Cluster                           | Effect Value                           |
| 1                                   | Hotelling’s Trace                      |
| 3.000                               | 4.703                                  |
| 2                                   |                                        |
| 35.000                              |                                        |

From the Table 4 we can see that it formed 2 clusters with the number of members in each clusters are 3 and 35, and Table 5 shows the T2 Hotelling value is 4.703 that transformed into statistics $F = 1.4805$. Compared to the value of $F_{0.1;1;36} = 2.8503$ it turns out that the three variables can not yet prove the difference of 2 formed clusters. Therefore, we go back with clustering process with form the higher number of clusters that is 3 clusters.
From the Table 6, we can see that it formed 3 clusters with the number of members in each cluster are 6, 29 and 3, and Table 7 we can conclude that with performing 3 clusters clustering process, there is only restlessness and percent rice consumption that have statistics $F > F_{0.1;2,35} = 2.46093$ that means there is only 2 variables have the differentiator power of 3 clusters clustering process. Therefore, we go back with clustering process with the higher number of clusters, that is 4 clusters.

From the Table 8, we can see that it formed 4 clusters with the number of members in each cluster are 2, 29, 1 and 6, and Table 9 we can conclude that with performing 4 clusters clustering process, there is only restlessness and percent rice consumption that have statistics $F > F_{0.1;3,34} = 2.2523$ that means there is only 2 variables have the differentiator power of 4 clusters clustering process. Therefore, we go back with clustering process with the higher number of clusters, that is 5 clusters.

From the Table 10, we can see that it formed 5 clusters with the number of members in each cluster are 1, 2, 6, 28, and 1, Table 11 we can conclude that with performing 5 clusters clustering process, there is only restlessness and percent rice consumption that have statistics $F > F_{0.1;4,33} = 2.1234$ that means there is only 2 variables have the differentiator power of 5 clusters clustering process. Therefore, we go back with clustering process with the higher number of clusters, that is 6 clusters.

From the Table 12, we can see that it formed 6 clusters with number of members in each cluster are 6, 22, 1, 7, 1, and 1, Table 13 we can conclude that with performing 6 clusters clustering process, all variables that are restlessness, Polity, and percent rice consumption that have statistics $F > F_{0.1;5,32} = 2.0360$ that means the three variables have the differentiator power of 6 clusters clustering process. Therefore, in this step clustering process stopped.

Further, evaluation of the criteria of success for population which is has similarities to Indonesian’s. From the result of clustering process with 6 clusters, it shows all criteria is fulfilled by...
cluster 2, that is there is Indonesia in the cluster, there are both crisis and non crisis countries in 2008, and it is shows the most number of cluster’s members compared to another clusters.

5. Conclusion and Suggestion

5.1. Conclusion
The conclusion of this study is there are 21 countries which is have similarities to Indonesian’s based on three characteristics (democratic atmosphere, rice consumption, and rice purchasing ability (restlessness)), that are Brazil, Costa Rica, Djibouti, Dominican Republic, Ecuador, Fiji, Guinea-Bissau, Haiti, India, Jamaica, Japan, South Korea, Madagascar, Malaysia, Mali, Nicaragua, Panama, Peru, Senegal, Sierra Leone and Suriname.

5.2. Suggestion
The suggestion based on this study result is so that this study can become preference for the next study which is perform the analysis to test the strength and predictability of indicator Unprecedented Restlessness (UR) as indicator of rice crisis.

References
[1] AFP. (2008, April 13). Haiti PM Ousted Over Soaring Food Prices.
[2] F.Hair, J., JR, A. R., Tatham, R. L., & Black, W. C. (1998). Multivariate Data Analysis. New Jersey: Prentice Hall International Inc.
[3] FAO. (2015). 1 Oktober 2015. www.faostat3.fao.org.
[4] FAO. (2010). The Rice Crisis: Markets, Policies, and Food Security. London: Earthscan.
[5] FAO. (2008). The State of Food Insecurity in the World 2008. Roma: FAO FIAT PANIS.
[6] Hajarisman, N. (2008). Statistika Multivariat. Bandung.
[7] Hidayat, Y. (2015). Developing Unprecendented Restlessness as the New Strong Indicator of Rice Crisis at National Level. Global Journal of Pure and Applied Mathematics.
[8] Hidayat, Y. (2013). Forecasting Framework of Rice Crises Time in Bandung-Indonesia.
[9] Johnson, R. A. (2002). Applied Multivariate Statistical Analysis. New Jersey: Prentice Hall.
[10] Marshall, M. G. (2014). Polity IV Project: Data User’s Manual. 16-17.
[11] Mattjik, A. A. (2011). Sidik Peubah Ganda. Bogor: Institut Pertanian Bogor Press.
[12] Mohanty, S. (2013). Trends in Global Rice Consumption.
[13] Mukhopadhyay, P. (2009). Multivariate Statistical Analysis. Singapore: World Scientific.
[14] Rencher, A. C. (2002). Methods of Multivariate Analysis. Canada: WILEYINTERSCIENCE.
[15] Rokach, O. M. (2005). Data Mining and Knowledge Discovery Handbook. Secaucus, NJ, USA: Springer-Verlag New York, Inc.
[16] Srihari, S. (2008). Principal Component Analysis & Multidimensional Scaling.