EASR

Engineering and Applied Science Research

https://www.tci-thaijo.org/index.php/easr/index

Published by the Faculty of Engineering, Khon Kaen University, Thailand

# Clustering countries according to the world happiness report 2019

M. Mujiya Ulkhaq*[1, 2] and Arga Adyatama[3]

[1]Department of Economics and Management, University of Brescia, Brescia 25121, Italy
[2]Department of Industrial Engineering, Diponegoro University, Semarang 50275, Indonesia
[3]PT Algoritma Data Indonesia, South Jakarta 12950, Indonesia

Received 27 April 2020
Revised 20 July 2020
Accepted 29 July 2020

**Abstract**

Since the first initiative in 2012, the World Happiness Report (WHR) has drawn international attention as it can help the policymakers to evaluate their policy options. In the 2019 edition, the WHR introduced six factors to describe the variation of happiness across the countries. Finland is declared as the happiest country just like what they did in the previous year. This study tried to cluster the countries according to the WHR 2019. Nine clustering algorithms were presented, and three internal validation indices were utilized to compare the algorithms. *k*-medoids algorithm was selected to illustrate how three distinguished clusters generated from the algorithm are different from each other. This study is expected to give an insight into how to implement clustering algorithms into the real-world data set and how to interpret the results.

**Keywords:** Clustering, Clustering algorithm, Cluster validation, The world happiness report

## 1. Introduction

The term happiness which refers to love, positive well-being, contentment, the experience of joy, awe, or pride, combined with a feeling that some one's life is worthwhile, meaningful, and good [1], is progressively a common subject in the cross-national study. It regards as an appropriate measurement for social growth and public policy goal. The first comparative investigation on happiness was arguably conducted by Cantril [2], in which fourteen countries as representative samples have been analyzed in the study. Since then, happiness has been embraced in several international survey programs, for instance, the World Value Survey, the Euro-barometer, the European Welfare Survey, and ultimately, the World Happiness Report (WHR).

The WHR is a landmark survey about the global happiness' state which positions countries around the globe by how happy people perceive themselves to be. The report was authored by independent professionals acting in their competences. The initiative began in 2012 as the motivation is to achieve policies to increase people's happiness in order to increase the public's national income. Since then, a series of reports continues to obtain global appreciation as academics, businesses, organizations, and civil society progressively employ happiness to inform their policy-making decisions.

The WHR 2019 [3] introduced six factors, i.e., gross domestic product (GDP) per capita, social support, healthy life expectancy (HLE), freedom to make life choices, perception of corruption, and generosity to explain the happiness. The report revealed that Nordic countries are among the top happiest countries. Specifically, Finland is on the top list with a total score of 7.769; followed by Denmark, Norway, and Iceland in the second, third, and fourth places, respectively; while Swedes lied on the seventh after Dutch and Swiss.

This research aims to cluster countries according to the WHR 2019. Clustering is a process of classifying objects, observations, or data which have feature(s) into groups (or clusters). Clustering has been addressed in many contexts and by researchers in many disciplines, such as in biology [4], marketing [5-7], psychology [8], image processing [9], and pattern recognition [10]. Many different types of clustering algorithms have been proposed in the literature. In this research, nine clustering algorithms are presented and then compared to look for "the best" way to cluster the countries. The WHR 2019 will be the basis information to perform clustering. As the best of our knowledge, such attempts are only presented in blogs or websites and never been brought into the academic field. Therefore, this study is expected to give an insight into how to implement clustering algorithms according to the WHR 2019 as well as how to interpret the results.

This research employed R, a programming language for statistical computing and graphics. It is motivated by the recognition of R in the field of statistics, data mining, and machine learning; and also, by the aid of its well-established clustering packages. This study is also intended to assist researchers who have programming skills in R language but have little experience in clustering data.

The paper is structured as follows. The following section presents the data set that was used in this research as well as the procedure of data cleansing and imputation. A concise overview of the clustering process is described in Section 3. Next, the results of each clustering algorithm and the performance evaluation to compare the algorithms are demonstrated. The interpretation of the selected algorithm's result also presented in this section. Finally, the conclusion and future research direction are presented in the last section.

*Corresponding author.
Email address: ulkhaq@live.undip.ac.id
doi: 10.14456/easr.2021.16

## 2. Data

### *2.1 Data set*

The data set used in this research is adopted from the online data collection reported on the website of the WHR (https://worldhappiness.report). Although the WHP 2020 [11] has been recently published in March 2020, the data set is not available for the public yet until this paper is written; hence, the 2019 data set is used instead. The WHR 2019 focused on the happiness and the community; specifically, it discussed how the happiness has evolved over the past several years as well as how the information and communication technology, social norms, and governance affect the communities.

The data set is a panel data which contains 25 columns and 1704 rows, omitting the first row as a heading row. The rows represent a total of 165 countries and which year the data is collected. The columns represent two descriptors, one response variable, six predictors of the response variables, and several additional variables which were calculated or collected from external sources. The descriptors are "`country name`", i.e., the name of the surveyed country and "`year`" as the year of data collection. The response variable is happiness score or subjective well-being (SWB) (later on we called this variable as "`life_ladder`"). The information of SWB was collected from the Gallup World Poll's (GWP) survey. It is the national average responses of the citizens to the following question (called the "Cantril ladder question"): "Visualize such a ladder which has number 0 at the bottom and 10 at the top of the ladder. The bottom is representing the worst case, i.e., the worst possible life for you, while the top is representing the best case. On which step of the ladder would you say you feel standing at this time?"

There are six proposed predictors of the response variable. The first is GDP per capita in purchasing power parity (PPP) at constant 2011 international dollars ("`log_gdp`"). The information was obtained from the World Bank's World Development Indicators (WDI). The data then were normalized by taking its natural logarithm. The second is HLE at birth ("`hle`"). The information was coming from the World Health Organization's (WHO) Global Health Observatory. The third predictor is social support ("`social_ support`"). This variable which is also from GWP survey is the national average responses of the respondents to the binary question as follows: "If you were in trouble or having a problem, do you have friends or relatives you were able to count on to assist you whenever you need them, or not?" The next predictor is freedom to make life choices ("`freedom`"). The variable is the national average of responses of the citizens for the GWP question as follows: "Do you feel satisfied or dissatisfied with your freedom to decide what you do in your life?" The fifth is generosity ("`generosity`"). It is the residual of regressing national average respondents' responses to the GWP question on GDP per capita. The corresponding GWP question is as follows: "Have you ever donated your money to a charity event in the past month?" The last predictor is perception of corruption ("`corruption`"). The information is obtained from the national average of the citizens' responses to two binary GWP questions as follows: "Is the corruption widespread in the government, or not?" and "Is the corruption also widespread in businesses, or not?"

We omitted the other variables since only the six predictors (`log_gdp`, `hle`, `social_support`, `freedom`, `generosity`, and `corruption`) are used to describe happiness (`life_ladder`) [3].

### *2.2 Data cleansing and imputation*

Since the WHR 2019 stated that only the average value of `life_ladder` from 2016 to 2018 was used so that only this particular range of years was included in this research. However, several countries have no information in one or more of the predictor variables. In this case, earlier information was used as if they are the 2016-2018 data. Three-years limit for how far back in looking for the missing values was applied. This filter made several countries being discarded to be analyzed further, i.e., Angola, Belize, Cuba, Djibouti, Guyana, Oman, Somaliland Region, Sudan, and Suriname.

Some information about the rest of the country was still missing since there is no such information in the original data set. For estimating `log_gdp` of Somalia and Taiwan, the most recent PPP statistics of GDP per capita from The World Factbook [12] was used. Swaziland, Taiwan, and the Palestinian Territories are still missing the information of `hle` so that the information from Salomon and his colleagues [13] was used. The missing values of `social_support`, `freedom`, and `generosity` of Qatar were estimated from the average of 2011-2012 data. Eight countries, i.e., Bahrain, Kuwait, China, Jordan, Taiwan, Turkmenistan, the United Arab Emirates, and Saudi Arabia, have missing values of `corruption`. For those countries, the missing information was estimated by using the "control of corruption" indicator from WGI [14]. A regression model containing WGI's control of corruption as a predictor and GWP's perception of corruption as a response variable was established. The missing values were then estimated by utilizing the established regression model. Finally, the statistics of Cyprus were used as information for Northern Cyprus' missing values.

## 3. Clustering: An overview

Clustering is regarded as one of the most useful methods in machine learning and data mining for finding the existence of groups (called clusters) as well as investigating interesting patterns in the data set. It is about dividing, separating, or partitioning the data set into clusters. The general objective is that the objects or data points or observations in a cluster are closer to (or more similar) each other than other data points in different clusters [15]. Clustering analysis can be applied in many disciplines, such as psychology, life sciences, marketing, engineering, and medical sciences. It might be found under different terms, for instance, typology (in social sciences), numerical taxonomy (in biology, ecology), unsupervised learning (in pattern recognition), and partition (in graph theory) [16]. There are no predefined groups or classes (called "the ground truth label") in the clustering analysis that show what kind of associations or relations among the data. For this reason, clustering analysis is also called unsupervised learning. Classification is a counterpart of clustering analysis as the predefined classes or categories are available (it is also called supervised learning).

The basic steps in the clustering process can be summarized as follows (see also Figure 1).

1. Data cleansing and imputation

   Real-world databases often contain errors (trivial or non-trivial, syntactic or semantic) and missing values. Data preprocessing might be necessary—to ensure the information is consistent, accurate, and high-quality—prior to their utilization in clustering analysis. Refer to the previous section to recall the process of data cleansing and imputation in this research.

2. Feature selection

   This step aims to choose proper features on which clustering analysis is to be conducted. Some works of literature relate this step with dimensionality reduction when it deals with high-dimensional data. Principle component analysis is typically used. It deals with constructing a linear combination of a set of vectors which could explain the data variance. However, since there are only plenty of features

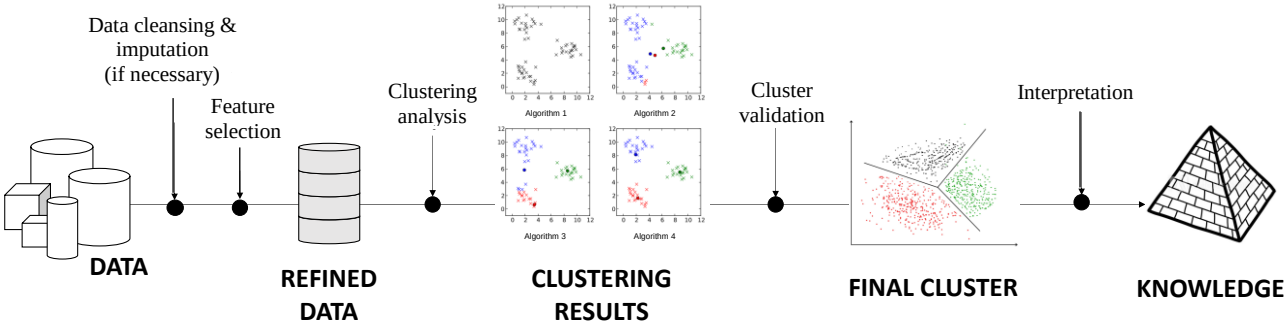



**Figure 1** Steps of the process of clustering

used in this research, this approach is not applied. Moreover, the result of the clustering analysis could be different with and without the dimensionality reduction [15]; also, the computational time is not a vital issue in this research. Rather, in this research, this step was performed by utilizing the multiple linear regression analysis (see Subsection 4.2). The predictors which significantly explain the response variable would be used as features for the next step.

3. Clustering analysis
   It refers to the choice of clustering algorithms. Several clustering algorithms have been proposed by scholars. Obviously, it is not possible to present and review all the algorithms; instead, in the following subsection, only algorithms used in this study will be presented.

4. Cluster validation
   Once clusters have been obtained by performing clustering algorithms, such a question could arise: "How well do the obtained clusters fit the data set?" The question is essential since several different clustering algorithms (or different configurations of similar clustering algorithm) could generate different clusters [17]; thus, one could analyze different clustering algorithms and choose the algorithm that best fits the data.

5. Interpretation
   In several cases, experts and professionals in the field of application somehow have to integrate the result obtained from the clustering algorithm with other analyses or experimental evidence to draw a correct conclusion as well as gain insightful knowledge.

### *3.1 Clustering algorithm*

Scholars have proposed several different types of clustering algorithms. In addition, several taxonomies to structure those different types of algorithms were available, see e.g., [18-21]. In this subsection, nine clustering algorithms used in this study are presented. The algorithms were selected according to the type of data used, the objective of the algorithms, as well as to represent each type of the clustering algorithm. The following is a brief explanation for each algorithm used in this research.

1. *k*-means
   *k*-means [22] is arguably the most broadly clustering algorithm used in literature due to the computational speed and its simplicity. It requires a distance matrix and a number of clusters (*k*). Initially, each object or observation is connected with one cluster according to its distance to the centroid or cluster center. The objective of this algorithm is to minimize the average squared distance between observations in the same cluster. The predefined number of clusters is one of the main limitations of this algorithm since the final clusters depend on the choice of the number of clusters. Moreover, *k*-means is considered as sensitive to the initial seed selection [21].

2. *k*-means++
   As previously stated, one of the drawbacks of *k*-means is that the algorithm is sensitive to the initialization of the centroids. In sum, a poor initialization could result in a poor clustering. To overcome the drawback, *k*-means++ by Arthur and Vassilvitskii [23] was proposed. This algorithm guarantees smarter initialization and improves the clustering quality. Apart from the initialization, the rest of the algorithm is similar to the standard *k*-means.

3. *k*-medoids
   This algorithm is also called as partitioning around medoid (PAM). It was proposed by Kaufman and Rousseeuw [24]. A medoid can be defined as the representative of the objects in the cluster, whose dissimilarities with all the other objects in the cluster is minimum. It is considered as a less sensitive (or robust) alternative to the *k*-means algorithm since *k*-medoids uses medoids as centroids as an alternative of means which is used in *k*-means.

4. Clustering large applications (CLARA)
   This algorithm [25] is an extension to *k*-medoids which deals with huge data (having more than several thousand objects or data points). This extension aims to reduce storage problems and computational time. Instead of identifying all medoids for all data set, the algorithm considers only a small sample of the data with a fixed size. Consequently, the *k*-medoids algorithm is applied to look for an optimal number of medoids for the predefined sample. CLARA repeats the sampling and clustering processes a pre-specified number of times to minimize sampling bias.

5. Agglomerative nesting (AGNES)
   Previously mentioned algorithms are the type of partitioning clustering. It means that they classify the observations into *k* clusters, in which each cluster has at least one observation and each observation belongs to exactly one cluster. In addition, two different clusters cannot have the same observation(s) and that the *k* clusters would add up to the full data set [24]. AGNES (and also the following algorithm, DIANA) belong to the type of hierarchical clustering. AGNES, in particular, starts by considering one observation as one cluster. Pairs of clusters are combined sequentially until all clusters were merged into one "big cluster" that contains the entire observations. The result of this algorithm is a dendrogram; it is a tree-based representation of the entire observations. The algorithm uses (dis)similarity between each pair of observations in the entire data set. Then, it uses linkage function to merge observations which are in close proximity to form the dendrogram. If one would create clusters, the cut-off point of the hierarchical tree should be determined.

6. Divisive analysis (DIANA)

DIANA is also the type of hierarchical clustering which is the inverse of AGNES. It starts by including all observations in one "big cluster". Iteratively, the most heterogeneous pairs of observations would be separated into two subsets. This step is repeated until all observations are located at their clusters. This algorithm poses computational problems: the first step involved considering all possible partitions into two subsets; this might be infeasible because of a huge number of combinations. Consequently, some scholars have restricted their attention to AGNES. (In the literature, DIANA has been largely ignored; as a matter of fact, when people discuss the hierarchical clustering, they often mean AGNES [24].)

7. Affinity propagation

One of the main limitations of the *k*-means algorithm and also other similar algorithms is the number of clusters and the initial set of points has to be defined at first. Affinity propagation proposed by Frey and Dueck [26], on the other hand, takes similarity between pairs of observations as input parameters and considers all observations as potential exemplars. Real-valued messages are exchanged between data points which is updated in response to the values from other pairs. This updating happens iteratively until convergence, at which point the final exemplars are chosen, and hence the final clustering is given. Each iteration has two message-passing steps:

(i) Calculating responsibilities $r(i,k)$. It reflects the accumulated evidence for how well-suited point $k$ is to serve as the exemplar for point $i$, taking into account other potential exemplars for point $i$. Responsibility is sent from data point $i$ to candidate exemplar point $k$.

(ii) Calculating availability $a(i,k)$. It reflects the accumulated evidence for how appropriate it would be for point $i$ to choose point $k$ as its exemplar, taking into account the support from other points that point $k$ should be an exemplar. Availability is sent from candidate exemplar point $k$ to point $i$.

The main drawback of this algorithm is its complexity, which makes this algorithm most appropriate for small to medium-sized datasets.

8. Spectral clustering

Traditional clustering algorithms like *k*-means and *k*-means++ use an elliptical or spherical metric to group observations; thus, they would not perform well if the clusters are non-convex. Spectral clustering, on the other hand, can be considered as a generalization of the traditional clustering algorithm which is intended for this kind of situation. The algorithm of the spectral clustering algorithm is as follows:

(i) The starting point is an $N \times N$ matrix of pairwise similarities $s_{ij}' \geq 0$ between each data point. The data points are then represented in an undirected similarity graph $G = <V, E>$, where the $N$ vertices $v_i$ represent the data points and the edges $e_i$ are weighted by $s_{ij}'$.

(ii) Cluster the data points by partitioning $G$ by its connected components. All symmetric nearest neighbors are connected with edges weighted with $s_{ij}'$ and points that are not nearest neighbors are not connected.

(iii) Compute the graph Laplacian $L$.

(iv) Compute the eigen-decomposition of $L$. Find the $m$ eigenvectors $\boldsymbol{Z}_{N\times m}$ corresponding to the $m$ smallest eigenvalues of $L$, ignoring the trivial constant eigenvector.

(v) Use a standard clustering algorithm to cluster the rows of $\boldsymbol{Z}$ (see [27] for more elaboration).

9. Density-based spatial clustering of applications with noise (DBSCAN)

This algorithm which was proposed by Ester and coauthors [28] views clusters as high-density areas separated by low-density areas. It is one of the most well-known density-based clustering algorithms. The essential component of this algorithm is the concept of core samples, i.e., samples in high density areas. The algorithm is based on the idea of clusters and noise. For each point in a cluster, the neighborhood of a given area must contain at least a minimum number of points. The algorithm works as follows:

(i) Find all the neighbor points within "eps" (i.e., the neighborhood around a data point) and identify the core points or visited with more than "MinPts" (i.e., minimum number of data points within eps radius) neighbors.

(ii) For each core point if it is not already assigned to a cluster, create a new cluster.

(iii) Find recursively all its density connected points and assign them to the same cluster as the core point. Two points *a* and *b* are said to be density connected if there exists a point *c* which has a sufficient number of points in its neighbors and both the points *a* and *b* are within the "eps" distance. This is a chaining process; so, if *b* is neighbor of *c*, *c* is neighbor of *d*, *d* is neighbor of *e*, which in turn is neighbor of *a* implies that *b* is neighbor of *a*.

(iv) Iterate through the remaining unvisited points in the data set. Those points that do not belong to any cluster are called noise.

### *3.2 Clustering performance evaluation*

Clustering algorithms deal with several parameters; frequently they have to deal with noisy, incomplete and sampled data, as well as run in high dimensional spaces; hence, their performance could differ for different types of data in different applications. The method for evaluating the performance of clustering algorithms is called cluster validation; it regards as one of the most central concerns in clustering analysis [29].

There are two criteria proposed for cluster validation, i.e., compactness (or cohesion) and separation. The former means that the member(s) of each cluster should be as close as possible to each other; and the latter implies that the clusters should be widely spaced. Validity measures used for assessing the performance of the algorithms with respect to those previous two criteria can be classified into relative, internal, and external validation [29].

Relative validation assesses the clustering by changing different parameter values for the same algorithm (for instance, changing the number of clusters $k$). It is commonly used for investigating the optimal number of clusters. Internal validation is according to the information inherent to the data set and assesses the quality of the clustering algorithm without any external information. Conversely, the external validation measures the similarity between the clustering algorithm's result and the "correct" clusters (or the ground truth label) of the data set. Since the ground truth label is unavailable (this study used the real data set, not artificial data set), only relative and internal validations were used.

In this study, the elbow method was used as a relative validation. It is performed by running the particular algorithm several times with a rising number of cluster $k$. Its sum of squared errors is then calculated and plotted against the number of clusters $k$. If the plot seems like an arm, then the "elbow" of the arm corresponds to the optimal number of clusters.

**Table 1** Summary of the data

| Variables | *Mean* | *Stdev.* | *Min.* | *Max.* |
|---|---|---|---|---|
| life_ladder | 5.406 | 1.116 | 3.075 | 7.769 |
| log_gdp | 9.238 | 1.234 | 6.480 | 11.693 |
| hle | 63.847 | 7.358 | 41.850 | 76.500 |
| social_support | 0.804 | 0.122 | 0.305 | 0.977 |
| freedom | 0.773 | 0.123 | 0.441 | 0.980 |
| generosity | -0.006 | 0.153 | -0.340 | 0.586 |
| corruption | 0.740 | 0.174 | 0.102 | 0.942 |

There are several internal validation indices in the literature, yet in this study, only three indices are used as follows, see [30] for a more comprehensive discussion. Before, let us denote some notations used. First, let us define a data set $\underline{X}$ as a set of N observations characterized as vectors in an *F*-dimensional space: $X = \{x_1, x_2, \dots, x_N\} \subseteq \Re^F$. A partition in $X$ is a set of disjoint groups that divides $X$ into $k$ clusters: $C = \{c_1, c_2, \dots, c_k\}$. The center of cluster *i* or centroid *i* $c_i$, is the mean vector $c_i$; and the centroid of the data set $\bar{X}$ is the mean vector of the whole data set. The Euclidean distance between objects $x_l$ and $x_p$ denotes as $d_E(x_l,x_p)$.

1. Silhouette index (SI)
   SI is a normalized summation-type index which can be calculated as [31]:

$$SI(C) = \frac{1}{N}\sum_{c_i \in C}\sum_{x_l \in c_i}\frac{b(x_l,c_i) - a(x_l,c_i)}{\max\{a(x_l,c_i), b(x_l,c_i)\}}, \quad (1)$$

   where $a(x_l,c_i) = 1/|c_i - 1|\Sigma_{x_p \epsilon c_i} d_E(x_l,x_p)$ and $b(x_l,c_i) = \min_{c_j \epsilon C \setminus c_i}\left\{1/|c_j|\Sigma_{x_p \epsilon c_i} d_E(x_l,x_p)\right\}$. The compactness is assessed according to the distance between all the observations in the same cluster, while the separation is according to the closest neighbor distance. The value of SI ranges from −1 to 1. If the value of SI is near to −1, it implies that the observation is closer to other clusters than its cluster; otherwise, if SI is near to 1, it implies that the average distance to the cluster to which it belongs is smaller than to any other cluster. The value around zero indicates overlapping clusters. The higher the SI value, the more separated and compact are the clusters.

2. Dunn's index (DI)
   DI can be defined as the ratio between the minimum distance between two clusters and the size of the largest intra-cluster distance. The index can be calculated as [32]:

$$DI(C) = \frac{\min_{c_i \in C}\left\{\min_{c_j \in C \setminus c_i}\left(\delta(c_i, c_j)\right)\right\}}{\max_{c_i \in C}\left(\Delta(c_i)\right)}, \quad (2)$$

   where $\delta(c_i,c_j) = \min_{x_l \epsilon c_i}\min_{x_p \epsilon c_j}\{d_E(x_l,x_p)\}$ and $\Delta(c_i) = \max_{x_l,x_p \epsilon c_i}\{d_E(x_l,x_p)\}$. The cohesion is estimated by the nearest neighbor distance while the separation is estimated by the maximum cluster diameter. When DI is having a high value, it indicates a well-separated and compact cluster.

3. Calinski-Harabasz' index (CHI)
   The cohesion is estimated according to the distances from the observation in a cluster to the centroid; while the separation is based on the distance from the centroid to the global centroid. The higher the index the better. The index can be defined as [33]:

$$CHI(C) = \frac{N-k}{k-1}\cdot\frac{\sum_{c_i \in C}|c_i| d_E\left(\overline{c_i}, \overline{X}\right)}{\sum_{c_i \in C}\sum_{x_l \in c_i} d_E\left(x_l, \overline{c_i}\right)}. \quad (3)$$

## 4. Results and discussion

### *4.1 Descriptive statistics*

The data used in this study consists of 156 objects, which is the country, one response variable (life_ladder), and six predictors (log_gdp, hle, social_support, freedom, generosity, and corruption). Table 1 shows the summary of the data.

The distribution of the Cantril ladder question's answers could give a portray to compare happiness levels as well as inequality across the countries. The global average is 5.406 (out of 10) and the standard deviation is 1.116. Syria became the country which has the lowest score of the Cantril ladder question (3.075) while Finland has the highest score (7.769). The scores fluctuated significantly among population-weighted regions, where the highest score is North America, Australia and New Zealand (NAAZ) region (7.177), followed by Western Europe (6.845), Latin America and the Caribbean (5.939), Central and Eastern Europe (5.765), East Asia (5.690), the Commonwealth of Independent States (CIS) (5.247), Southeast Asia (5.334), the Middle East and North Africa (MENA) (5.216), South Asia (4.548), and Sub-Saharan Africa (4.303), see Figure 2. The happiness inequality can be evaluated by the standard deviation of the distributions of individual happiness scores. The lowest scores are found in Western Europe, NAAZ, and South Asia; while the largest scores are found in Latin America and the Caribbean, Sub-Saharan Africa, and MENA.

The country with the highest value of GDP is Qatar while Central African Republic has the lowest one. NAAZ is the region in which the highest average GDP and Sub-Saharan African region has the lowest value. The inequality in terms of HLE is very big as it shows by its standard deviation: the maximum value is 76.5 (Singapore) and the minimum value is 41.850 (Swaziland). The region with the highest value of HLE is Western Europe (72.690). Central African Republic has the lowest score of the GWP question about social support while Icelanders are among people who confidently answered that they do have relatives or friends whenever they are in trouble. Interestingly, the region which has the lowest value of freedom, meaning that people there are not really satisfied with their freedom to choose what they do with their lives. Another interesting fact is that Myanmar and Indonesian people are among the most generous people compared to other citizens, making Southeast Asia region is placed in the second position (after NAAZ region) for the highest generosity aspect. Singaporeans do believe that corruption is not widespread throughout both the government and within the business (the lower score is the better), while Moldovans are less confident that their government (and in the business as well) is not being corrupted. Central and Eastern Europe have the highest score

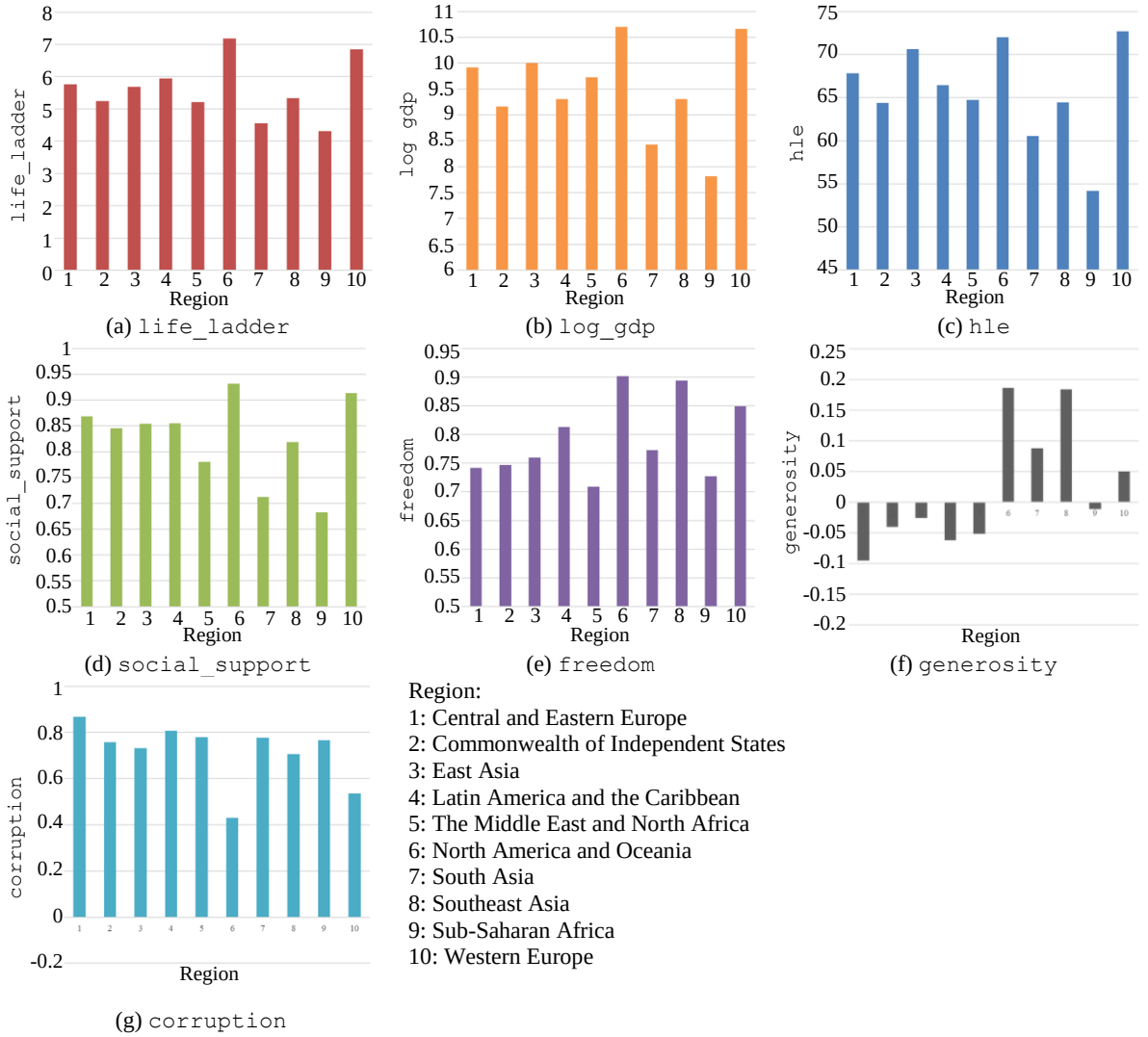



**Figure 2** The average value of each variable for each region

**Table 2** The regression result

| Variables | *Original Model* | | *Refined Model* | |
|---|---|---|---|---|
| | ***Estimates*** | ***p-value*** | ***Estimates*** | ***p-value*** |
| constant | -2.355 | .000** | -2.335 | .000** |
| log_gdp | .255 | .000** | .250 | .000** |
| hle | .033 | .003** | .034 | .003** |
| social_support | 2.741 | .000** | 2.741 | .000** |
| freedom | 1.892 | .000** | 1.925 | .000** |
| generosity | .125 | .681 | - | - |
| corruption | -.524 | .071* | -.552 | .050* |

**Coefficient is significant at the 0.05 level (2-tailed)
*Coefficient is significant at the 0.10 level (2-tailed)

while North America and Oceania are having the lowest score for this predictor.

*4.2 Feature selection*

The multiple linear regression model was used to describe how six predictors (log_gdp, hle, social_support, freedom, generosity, and corruption) explain the variation of life_ ladder across countries.

The estimates of the regression parameter model are shown in Table 2 (see the second column). The third column denotes the *p*-value for the hypotheses test of the population regression parameters. The asterisk symbol * indicates that the test is significant at $\alpha$ = 10% whereas a double asterisk ** denotes that the test is significant at $\alpha$ = 5%. Note that the coefficient for corruption is statistically significant at the level of 10%, the coefficient for generosity is not statistically significant, and the rest are significant at the 5% level.

Taking all together, the six predictors explain more than 77% of the variation in happiness as a response variable, among countries being investigated. Specifically, the sample coefficient of determination $R^2$ is 78.0%, while the adjusted $R^2$ is 77.1%. (Theil [34] suggested to use adjusted $R^2$ than $R^2$ since $R^2$ is likely to yield an overly optimistic image of the fitted value, especially when the number of predictors is not too small compared to the number of observations.) The value tells the proportion of variation in the response variable described by the predictor variables. In this case, the value of adjusted $R^2$ equals to 77.1%

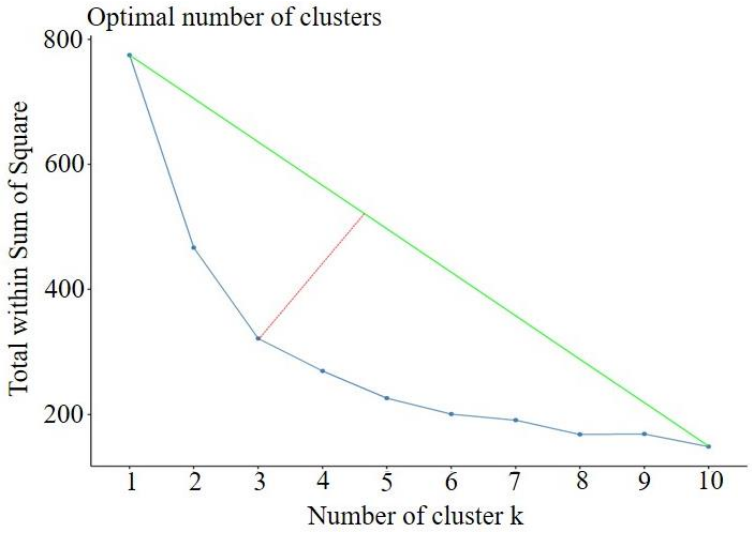



**Figure 3** The elbow graph for the *k*-means algorithm

means that 77.1% of the variability of SWB can be described by the previously mentioned regression model. This value is a sign of a good model.

Since the variable generosity is not statistically significant, this variable is discarded for the next analysis. The new linear regression model which only consists of five predictors have all statistically significant coefficients (see the last column of Table 2) while the adjusted $R^2$ increases to 77.2%. The overall model (by employing the analysis of variance) is also statistically significant (*p*-value = 0.000). Finally, to validate the model, the classical assumptions also were checked. Note that there is no evidence for autocorrelation among the errors; no evidence that a predictor is collinear with the other predictors; and lastly, heteroscedasticity does not present.

*4.3 Clustering results*

Before conducting clustering analysis with several clustering algorithms, it is necessary to transform the data into a standardized value (or z-score) [35], which has mean equals to 0 and standard deviation equals to 1. The standardized value can be defined by:

$$Z_{ij} = \frac{X_{ij} - \overline{X_j}}{s_j}, \quad (4)$$

where $Z_{ij}$ is the *z*-score for object *i* and variable *j*, $X_{ij}$ is the original data value, $\overline{X_j}$ is the mean or average of variable *j*, and $s_j$ is the standard deviation of variable *j*. This standardized value has an advantage of unitless (the numerator and the denominator are in the same units [24]). In addition, it is beneficial for the next analysis.

The first algorithm used is *k*-means. kmeans script (in R) is used; it is the function in stats package. In R, the format is kmeans(x, centers), where x is the observations and centers is the predefined number of clusters. In this study, the elbow method was used to investigate the optimal number of clusters (see Subsection 3.2). The elbow graph is depicted in Figure 3. Note that the curve is plotted in the blue color, while the green line connects the start and end points of the curve, and the red line is orthogonal to the green line that crosses the blue curve, maximizing the distance between the red line and the blue curve. It gives the optimal number of clusters = 3. The result is displayed in the second column in the Appendix. The first cluster has 87 members (countries), the second cluster has 46 members, and cluster number three has only 23 members. The centroid (the mean) of each cluster is shown in Table 3.

The second algorithm is *k*-means++. `kmeanspp` script is used; it is the function in `LICORS` package. The format is `kmeanspp(x,centers)`. The result is shown in the Appendix (the third column), while the centroid of each cluster is shown in Table 3. Note that the result is identical to *k*-means, only the "labeling" is different: the second cluster in *k*-means is cluster number three in *k*-means++; vice versa. It is a normal condition in the clustering algorithm since the label is chosen arbitrarily.

Next, *k*-medoids (PAM) is utilized under the function pam in `cluster` package. The format is `pam(x,centers,metric)`, where metric specifies the distance metrics to be used ("`metric=eucledian`" was used, meaning that we used the Euclidean distance). By also employing the elbow graph, the optimal number of clusters is found, i.e., 3. The result is shown in the Appendix (the fourth column), while the medoid of each cluster is presented in Table 3. Note that the cluster membership is almost similar to *k*-means (or *k*-means++), yet only three countries, i.e., Estonia, Morocco, and the United Arab Emirates, have different memberships.

The function `clara` in `cluster` package is used for identifying cluster membership in the CLARA algorithm. The format is `clara(x,centers,metric,samples)`, where the Euclidean distance was also used and `samples` means the number of samples drawn from the data set ("`samples=50`" was chosen). By also employing the elbow graph, the optimal number of clusters is identified, i.e., 3. The result is shown in the Appendix (the fifth column). It is of interest to see that the cluster membership is identical to *k*-medoids, even also the label. The result is unsurprising since the data used is not considered large enough so that the algorithm behaves like PAM.

Affinity propagation algorithm can be executed by using `apcluster` function in `apcluster` package. The format is `apcluster(s,x)`, where `s` is similarity matrix or similarity function. The `negDistMat` with `r=2` from `Matrix` package was used as a similarity function. It created a square matrix of mutual pairwise similarities of vectors as negative distances (`r=2` is applied to obtain negative squared distances as what Frey and Dueck demonstrated, the ones who proposed the algorithm). The algorithm results 13 clusters (see the sixth column in the Appendix). To do the spectral clustering algorithm in R, `specc` function from `kernlab` package is used. The format is `specc(x,centers)`, where `centers=3` was used. The result of the spectral clustering is depicted in the Appendix of the

**Table 3** The cluster centers of each algorithm's result

| Clusters (number of members) | log_gdp | hle | social_support | freedom | corruption |
|---|---|---|---|---|---|
| ***k*-means:** | | | | | |
| Cluster 1 (87) | 9.626 | 66.421 | 0.850 | 0.773 | 0.817 |
| Cluster 2 (46) | 7.770 | 54.819 | 0.656 | 0.708 | 0.755 |
| Cluster 3 (23) | 10.707 | 72.164 | 0.927 | 0.905 | 0.419 |
| ***k*-means++:** | | | | | |
| Cluster 1 (87) | 9.626 | 66.421 | 0.850 | 0.773 | 0.817 |
| Cluster 2 (23) | 10.707 | 72.164 | 0.927 | 0.905 | 0.419 |
| Cluster 3 (46) | 7.770 | 54.819 | 0.656 | 0.708 | 0.755 |
| ***k*-medoids (PAM):** | | | | | |
| Cluster 1 (45) | 7.594 | 53.3 | 0.647 | 0.732 | 0.777 |
| Cluster 2 (90) | 9.762 | 68.0 | 0.850 | 0.810 | 0.806 |
| Cluster 3 (21) | 10.787 | 72.2 | 0.934 | 0.916 | 0.389 |
| **CLARA:** | | | | | |
| Cluster 1 (45) | 7.594 | 53.3 | 0.647 | 0.732 | 0.777 |
| Cluster 2 (90) | 9.762 | 68.0 | 0.850 | 0.810 | 0.806 |
| Cluster 3 (21) | 10.787 | 72.2 | 0.934 | 0.916 | 0.389 |
| **Affinity propagation:** | | | | | |
| Cluster 1 (13) | 10.587 | 71.982 | 0.903 | 0.865 | 0.588 |
| Cluster 2 (14) | 8.972 | 61.168 | 0.827 | 0.832 | 0.698 |
| Cluster 3 (1) | 6.480 | 45.050 | 0.305 | 0.635 | 0.874 |
| Cluster 4 (15) | 7.709 | 52.738 | 0.636 | 0.739 | 0.756 |
| Cluster 5 (8) | 7.498 | 55.136 | 0.601 | 0.496 | 0.774 |
| Cluster 6 (13) | 9.915 | 67.038 | 0.904 | 0.670 | 0.841 |
| Cluster 7 (16) | 9.434 | 65.760 | 0.780 | 0.619 | 0.854 |
| Cluster 8 (6) | 9.045 | 64.650 | 0.635 | 0.794 | 0.750 |
| Cluster 9 (22) | 10.313 | 68.732 | 0.885 | 0.831 | 0.823 |
| Cluster 10 (2) | 7.419 | 55.200 | 0.600 | 0.917 | 0.307 |
| Cluster 11 (20) | 9.112 | 65.871 | 0.838 | 0.862 | 0.844 |
| Cluster 12 (12) | 10.896 | 73.042 | 0.938 | 0.925 | 0.291 |
| Cluster 13 (14) | 7.641 | 54.482 | 0.744 | 0.723 | 0.815 |
| **Spectral Clustering:** | | | | | |
| Cluster 1 (84) | 9.724 | 66.734 | 0.854 | 0.778 | 0.817 |
| Cluster 2 (21) | 10.707 | 72.594 | 0.930 | 0.905 | 0.396 |
| Cluster 3 (51) | 7.833 | 55.488 | 0.669 | 0.712 | 0.754 |
| **AGNES:** | | | | | |
| Cluster 1 (48) | 7.825 | 55.322 | 0.658 | 0.714 | 0.753 |
| Cluster 2 (81) | 9.562 | 66.119 | 0.853 | 0.769 | 0.823 |
| Cluster 3 (27) | 10.777 | 72.184 | 0.916 | 0.892 | 0.467 |
| **DIANA:** | | | | | |
| Cluster 1 (53) | 7.865 | 55.605 | 0.674 | 0.719 | 0.757 |
| Cluster 2 (78) | 9.712 | 66.788 | 0.930 | 0.769 | 0.825 |
| Cluster 3 (25) | 10.670 | 72.143 | 0.669 | 0.901 | 0.439 |

seventh column. The centroids of the algorithm are presented in Table 3.

The only density-typed clustering used in this study, i.e., DBSCAN, is run by utilizing function `dbscan` in `dbscan` package. The format is `dbscan(x,eps,MinPts)`, where `eps` is the size of the epsilon neighborhood and `MinPts` is the number of minimum points in the eps region. To determine the `eps` value, the `kNNdist` function was used: `kNNdist(x,k=4)`. The idea is to calculate the average of the distances of every object to its k closest neighbors. Next, these *k*-distances would be plotted in ascending order. The aim is to define the "knee", which corresponds to the optimal `eps`. A knee is defined as a threshold where a sharp change occurs along the *k*-distance curve. The curve is depicted in Figure 4. One can observe that the optimal eps is around 1.4. Note that in DBSCAN, there are no cluster centers, and clusters are produced by linking adjacent points to one another. The result of the algorithm is depicted in the Appendix of the eighth column. There is only one cluster, while four countries (Central African Republic, Rwanda, Somalia, and Uzbekistan) are called "noises" that do not belong to any cluster.

The last two algorithms, AGNES and DIANA, which belong to hierarchical clustering would be analyzed differently. Those two can be executed by using `agnes` and `diana` functions in `cluster` package. The format is `agnes`—or `diana`—`(x,diss,method, metric,stand)`, where `diss` is a logical flag: if `T` (or true) is chosen, then `x` is treated as it is a dissimilarity matrix, otherwise, `x` is assumed to be a matrix of observations by variables (`diss=F` was used in this study). This research used `ward` (from Ward's method) in the argument `method`, which minimizes the total within-cluster variance. It implies that at each iteration, the pair of clusters with minimum between-cluster distance will be combined. The Euclidean distance also used to fill the argument `metric`. The argument stand is also a logical flag: if `T` is chosen, then `x` will be standardized first before the dissimilarities are computed; since `x` is already standardized, so `stand=F` was used.

The results of those two algorithms are dendrograms. The dendrogram is interpreted as follows. As we move up the dendrogram (or the hierarchical tree), similar countries or objects are merged into "a twig". Again, similar "combined objects" would be merged into a bigger twig (or branch). The process is

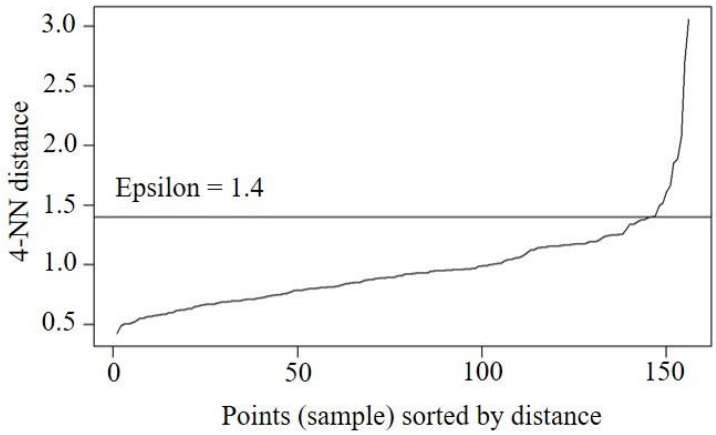
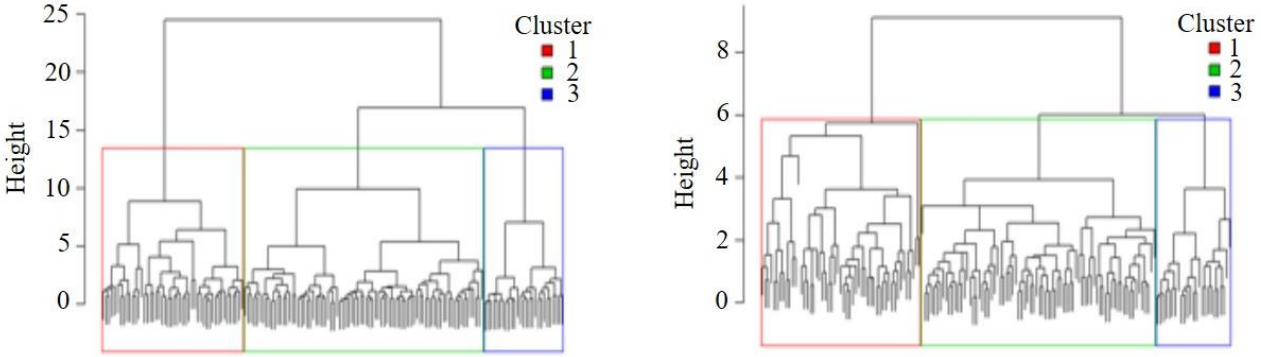



**Figure 4** *k*-distance curve




(a) AGNES (b) DIANA

**Figure 5** "Cut" dendrograms

repeated until we have one "completed tree" combining from bigger branches. The "height" of the fusion of the similar (combined) objects which is displayed on the vertical axis, shows the distance or (dis)similarity between two objects (or combined objects). The higher the height, the less similar the objects (or combined objects) are. To create partitions, one could "cut" the dendrogram at a certain value of height to divide the objects into clusters. The number of clusters equals three is chosen arbitrarily, just to show how the dendrograms are cut, see Figure 5. (It is hard to see the members since there are many countries.) The cluster membership when the dendrograms are cut to give three clusters is shown in the Appendix (ninth column for AGNES and tenth column for DIANA). For AGNES, the first cluster has 48 members, the second cluster has 81 members, and the rests belong to the third cluster. For DIANA, 53 countries belong to cluster number one, 78 countries belong to cluster number two, and cluster number three contains 25 countries. The centroids of each cluster are shown in Table 3.

*4.4 Cluster validation*

The previous subsection has demonstrated the algorithms as well as the results generated by each algorithm. This subsection would describe how to compare the algorithms based on the cluster validation techniques. The relative cluster validation using the elbow method to choose the optimal number of clusters. The external cluster validation cannot be used in this study since obviously, it is only able to be used in a controlled test environment. This study used real data set so that the structure of the data is unknown and hence, the correct cluster (or the ground truth) is unavailable [30]. Therefore, only the internal validation will be discussed here.

Table 4 shows the algorithms, along with three internal validation indices, i.e., SI, DI, and CHI, as well as the average within error sum of squares (SSE) and the average between SEE. Notice that two algorithms are not included in here, i.e., affinity propagation (since it might be different in many aspects with other algorithms) and DBSCAN (since it only contains one cluster so that it is not possible to calculate the distance between clusters). Spectral clustering has the minimum average within SSE while DIANA has the minimum average between SSE. Among the three internal validation indices, the algorithms being consideration are overlapping each other. According to SI, *k*-medoids and CLARA are considered as "the best" algorithm, while based on DI, AGNES is "the best"; however, *k*-means and *k*-means++ are "the best" according to CHI.

These results are not unanticipated since the differences in the previous internal validation indices make it hard to compare in the same environment. Some scholars showed that there is no single internal validation index which surpasses other indices [30], [36-39]. Therefore, it is not recommended to proclaim "the best" algorithm when comparing clustering algorithms [19]. The study from Arbelaitz and his colleagues [30] that was claimed to be the most extensive cluster validation index comparison ever published, compared 30 internal cluster validation indices in many different environments. The result of this study identified that SI produced the best results in most cases. Therefore, for the next discussion, the algorithm which has the highest SI value would be analyzed. There are two algorithms in this case, and *k*-medoids algorithm is selected arbitrarily.

*4.5 Interpretation of selected algorithm*

Providing users with meaningful insights from the original data could be considered as the ultimate goal of clustering analysis. It allows users to effectively solve the problems they face. This subsection would discuss how to interpret the algorithm result as we can gain some insights and knowledge.

**Table 4** Comparing clustering algorithms

| Algorithms | *Number of clusters* | *Average within SSE* | *Average between SSE* | *SI* | *DI* | *CHI* |
|---|---|---|---|---|---|---|
| *k*-means | 3 | 1.8872 | 3.5761 | 0.3693 | 0.0731 | 107.9154 |
| *k*-means++ | 3 | 1.8872 | 3.5761 | 0.3693 | 0.0731 | 107.9154 |
| *k*-medoids (PAM) | 3 | 1.8877 | 3.6111 | 0.3756 | 0.0916 | 107.6489 |
| CLARA | 3 | 1.8877 | 3.6111 | 0.3756 | 0.0916 | 107.6489 |
| Spectral Clustering | 3 | 1.8855 | 3.5560 | 0.3696 | 0.0916 | 107.0821 |
| AGNES | 3* | 1.9089 | 3.4962 | 0.3493 | 0.1326 | 102.2002 |
| DIANA | 3* | 1.8894 | 3.4935 | 0.3551 | 0.0876 | 106.0936 |

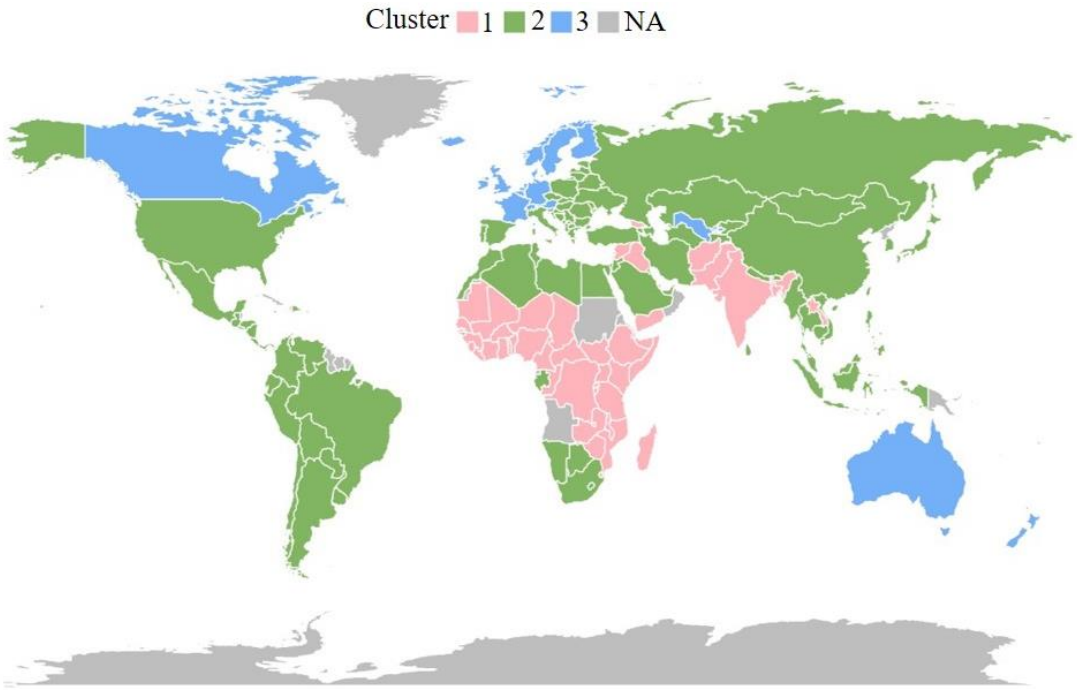


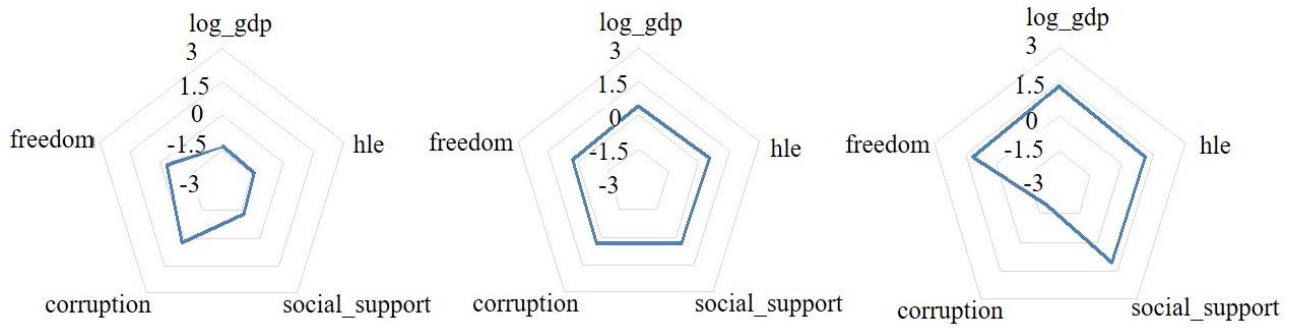

**Figure 6** Map of cluster membership according to the selected algorithm

Previously, the *k*-medoids algorithm was chosen to be analyzed. Note that it does not make *k*-medoids is the best algorithm among others since the decision is rather arbitrarily. The cluster membership is shown in the Appendix in the fourth column, while the map of cluster membership is shown in Figure 6. Each cluster's characteristics will be discussed as follows.

The first cluster consists of 45 countries, from six regions: four South Asian countries, one from Southeast Asia, CIS, and Latin America and Caribbean, three from MENA, and the rest are from Sub-Saharan Africa. The radar chart is depicted in Figure 7, where the values are the (standardized) medoids of each predictor variable, making it easier to compare. This cluster is having the lowest average value of GDP (the lowest GDP among all countries is in this cluster). In terms of HLE, this cluster is also the worst with the average value of only 54 years. The condition also happens in the other two predictors, i.e., social support and freedom to make life choices. The average values of social support (0.657) and freedom (0.706) are the worst among other clusters. An interesting fact is that the citizens of the countries belong to the first cluster perceive better in terms of perception of corruption than the second cluster. Comparing with the data from United Nations [40], all countries belong to low-income economies, i.e., which have gross national income (GNI) per capita less than 1,025 USD (except Nepal and Tajikistan), also the member of the first cluster. It is arguably to say that this first cluster has the least happy citizens.

There are 90 countries in the second cluster, making it the most widely spread cluster since the members are coming from all regions. This cluster has the "average" value (i.e., the medoids are around zero points) for all predictor variables. The members of the cluster have the average log GDP value of 9.642, making it less than the third cluster; despite the fact that Qatar as the country that has the highest GDP belongs to this cluster. The average value of HLE at birth is 66.439 years, also less than the third cluster. The average values of social support and freedom to make life choices are also in the second rank, below the third cluster. The government and the business operated in the countries of this cluster are perceived bad (the worst among all clusters) by the citizens in terms of corruption. Countries in Eastern Europe that have a bad reputation in this aspect belong to this cluster; the bottom four of this feature, i.e., Moldova, Romania, Bulgaria, and Bosnia and Herzegovina are the members of this cluster. Another interesting fact is that this cluster has the United States of America, the only country in the region of NAAZ. This country suffers from the bad perception of corruption and health facilities.

The last cluster, i.e., the third cluster, has all features that make it the happiest cluster. Its average values of all features are the best among all clusters. It contains only 21 countries; one from CIS (Uzbekistan), East Asia (Hong Kong), and Southeast Asia (Singapore), three from NAAZ (Canada, Australia, and New Zealand), and the rest are from Western Europe. The lowest value in HLE of this cluster (i.e., Uzbekistan with 64.8 years) is still better than the country which has the highest HLE in the first cluster (i.e., Bangladesh with 63.8 years). This situation also happens in social support feature, whereas the lowest value in this cluster (i.e., Hong Kong), is still better than the best country in terms of social support in the first cluster (i.e., Mauritania). Apart from the absence of the United States in this cluster (also other high-income countries, such as South Korea and Japan), another




(a) First cluster (b) Second cluster (c) Third cluster

**Figure 7** The radar charts of each cluster

surprising information is that Uzbekistan, a lower-middle-income country according to United Nations [40], is the member of this cluster. This country has the highest value of freedom to make life choices and also has high value of social support, compensating its low GDP and HLE.

## 5. Conclusion and future research direction

This research has demonstrated several clustering algorithms to cluster countries according to the WHR 2019. The features used as basis for clustering are `log_gdp`, `hle`, `social_support`, `freedom`, and `corruption`. Nine clustering algorithms were selected according to the type of the data used (all the numerical information are interval-type of data), the objective of the algorithms, and to represent each type of the clustering algorithm (i.e., partitioning clustering: *k*-means, *k*-means++, *k*-medoids; affinity propagation; spectral clustering; density-type clustering: DBSCAN; and hierarchical clustering: AGNES and DIANA). The result of each algorithm, i.e., the cluster membership, is depicted in the Appendix. Note that there is no "the best" clustering algorithm. *k*-medoids (PAM) that was selected as a representative of the algorithm to show the interpretation of its result was rather chosen arbitrarily. The (selected) final clustering contains three clusters whose characteristics are shown in Subsection 4.5. It is arguably to infer that the first cluster has the least happy citizens, while the third cluster is the happiest cluster.

When the data set of WHR 2020 is available to the public, it is demanding to adopt similar research and compare its result with this study. The "movement" of the country from its previous cluster to the "future" cluster (if any) will be an interesting discussion to follow. Another consideration is that since this study only takes the "hard" clustering algorithms into account, it is of interest to also apply the "fuzzy" clustering algorithm to richen the finding of the upcoming research. Lastly, this study only somehow connects the finding with the economic state of the country (i.e., the GNI). There are several recent works of literature that connected the happiness to sustainability, e.g., [41-43]; therefore, linking the country's cluster membership to its state of sustainability is an interesting area to be pursued.

## 6. References

[1] Lyubomirsky S. The how of happiness: a scientific approach to getting the life you want. New York: Penguin Press; 2008.

[2] Cantril H. The pattern of human concern. New Brunswick: Rutdgers University Press; 1965.

[3] Helliwell JF, Layard R, Sachs J. World happiness report 2019. New York: Sustainable Development Solutions Network; 2019.

[4] Kapourani CA, Sanguinetti G. Melissa: Bayesian clustering and imputation of single-cell methylomes. Genome Biol. 2019;20:61.

[5] Utami AA, Ginanjar AR, Fadlia N, Lubis IA, Ulkhaq MM. Using shopping and time attitudes to cluster food shoppers: An empirical finding from Indonesia. J Phys: Conf Ser. 2019;1284:012005.

[6] Minako FS, Ulkhaq MM, 'Sa Nu D, Pratiwi ARA, Akshinta PY. Clustering internet shoppers: An empirical finding from Indonesia. Proceedings of the 2019 5th International Conference on E-business and Mobile Commerce; 2019 May 22-24; Taichung, Taiwan. Taiwan: Association for Computing Machinery; 2019. p. 35-9.

[7] Ulkhaq MM, Fidiyanti F, Adyatama A, Maulani ZA, Nugroho AS. Segmentation of cinema audiences: an empirical finding from Indonesia. Proceedings of the 2019 2nd International Conference on Data Storage and Data Engineering; 2019 Jun 15-18; Jeju, Korea. Korea: Association for Computing Machinery; 2019. p. 3-8.

[8] Brusco MJ, Steinley D, Stevens J, Cradit JD. Affinity propagation: an exemplar-based tool for clustering in psychological research. Br J Math Stat Psychol. 2019;72:155-82.

[9] John R, Ramesh H. Colour based segmentation of a landsat image using k-means clustering algorithm. J Image Process Pattern Recogn Progr. 2017;4(3):31-8.

[10] Unglert K, Radić V, Jellinek AM. Principal component analysis vs. self-organizing maps combined with hierarchical clustering for pattern recognition in volcano seismic spectra. J Volcanol Geoth Res. 2016;320:58-74.

[11] Helliwell JF, Layard R, Sachs J, De Neve JE. World happiness report 2020. New York: Sustainable Development Solutions Network; 2020.

[12] The World Factbook [Internet]. 2020 [cited 2020 Apr]. Available from: https://www.cia.gov/library/publications/the-world-factbook/fields/208rank.html.

[13] Salomon JA, Wang H, Freeman MK, Vos T, Flaxman AD, Lopez AD, et al. Healthy life expectancy for 187 countries, 1990–2010: a systematic analysis for the Global Burden Disease Study 2010. Lancet. 2012;380:2144-62.

[14] The Worldwide Governance Indicators Project [Internet]. 2020 [cited 2020 Apr]. Available from: https://info.worldbank.org/governance/wgi/.

[15] Manly BFJ, Alberto JAN. Multivariate statistical methods: a primer. 4th ed. Boca Raton: CRC Press; 2017.

[16] Theodoridis S, Koutroumbas, K. Pattern recognition. 2nd ed. San Diego: Academic Press; 2008.

[17] Pal NR, Biswas J. Cluster validation using graph theoretic concepts. Pattern Recogn. 1997;30(6):847-57.

[18] Fraley C, Raftery AE. How many clusters? Which clustering method? answers via model-based cluster analysis. The Comput J. 1998;41(8):578-88.

[19] Xu R, Wunsch II D. Survey of clustering algorithms. IEEE Trans Neural Network. 2005;16(3):645-78.

[20] Mouton JP, Ferreira M, Helberg ASJ. A comparison of clustering algorithms for automatic modulation classification. Expert Syst Appl. 2020;151:113317.

[21] Jain AK, Murty MN, Flynn PJ. Data clustering: a review. ACM Comput Surv. 1999;31:264-323.
[22] McQueen J. Some methods for classification and analysis of multivariate observations. In: Le Cam LM, Neyman J, editors. Berkeley Symposium on Mathematical Statistics and Probability; 1965 Dec 27 – 1966 Jan 7; Berkeley, USA. USA: The Regents of the University of California; 1967. p. 281-97.
[23] Arthur D, Vassilvitskii S. k-means++: the advantages of careful seeding. Proceedings of the 18th annual ACM-SIAM symposium on Discrete algorithms; 2007 Jan 7-9; New Orleans, Louisiana. USA: Society for Industrial and Applied Mathematics; 2007. p. 1027-35.
[24] Kaufman L, Rousseeuw PJ. Finding groups in data: an introduction to cluster analysis. Hoboken: John Wiley & Sons; 1990.
[25] Kaufman L, Rousseeuw PJ. Clustering large data sets (with discussion). In: Gelsema ES, Kanal LN, editors. Pattern recognition in practice II. Amsterdam: Elsevier; 1986. p. 425-37.
[26] Frey BJ, Dueck D. Clustering by passing messages between data points. Science. 2007;315:972-6.
[27] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. 2nd ed. Canada: Springer; 2017.
[28] Ester M, Kriegel HP, Sander J, Xu X. Density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining. USA: AAAI; 1996. p. 226-31.
[29] Halkidi M, Batistakis Y, Vazirgiannis M. On clustering validation techniques. J Intell Inform Syst. 2001;17: 107-45.
[30] Arbelaitz O, Gurrutxaga I, Muguerza J, Pérez JM, Perona I. An extensive comparative study of cluster validity indices. Pattern Recogn. 2013;46(1):243-56.
[31] Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math. 1986;20:53-65.
[32] Dunn JC. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. J Cybernetics. 1973;3(3):32-57.
[33] Caliński T, Harabasz J. A dendrite method for cluster analysis. Commun Stat. 1974;3:1-27.
[34] Theil H. Introduction to econometrics. Prentice Hall: Englewood Cliffs; 1978.
[35] James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning with applications in R. 7th ed. New York: Springer; 2013.
[36] Dimitriadou E, Dolñicar S, Weingessel A. An examination of indexes for determining the number of clusters in binary data sets. Psychometrika. 2002;67:137-59.
[37] Maulik U, Bandyopadhyay S. Performance evaluation of some clustering algorithms and validity indices. IEEE Trans Pattern Anal Mach Intell. 2002;24(12):1650-4.
[38] Milligan GW, Cooper MC. An examination of procedures for determining the number of clusters in a data set. Psychometrika. 1985;50:159-79.
[39] Brun M, Sima C, Hua J, Lowey J, Carroll B, Suh E, et al. Model-based evaluation of clustering validation measures. Pattern Recogn. 2007;40(3):807-24.
[40] United Nations. World economic situation and prospects 2020. New York: United Nations; 2020.
[41] Cloutier S, Jambeck J, Scott N. The Sustainable Neighborhoods for Happiness Index (SNHI): a metric for assessing a community's sustainability and potential influence on happiness. Ecol Indicat. 2014;40:147-52.
[42] Cloutier S, Pfeiffer D. Sustainability through happiness: a framework for sustainable development. Sust Dev. 2015;23:317-27.
[43] Carlsen L. Happiness as a sustainability factor. The world happiness index: a posetic-based data analysis. Sustain Sci. 2018;13:549-71.

**Appendix** Cluster membership for each algorithm

| Country | *k*-means | *k*-mean++ | *k*-medoids (PAM) | CLARA | Affinity propagation | Spectral clustering | DBSCAN | AGNES | DIANA |
|---|---|---|---|---|---|---|---|---|---|
| Afghanistan | 2 | 3 | 1 | 1 | 5 | 3 | 1 | 1 | 1 |
| Albania | 1 | 1 | 2 | 2 | 8 | 1 | 1 | 1 | 2 |
| Algeria | 1 | 1 | 2 | 2 | 7 | 1 | 1 | 2 | 2 |
| Argentina | 1 | 1 | 2 | 2 | 9 | 1 | 1 | 2 | 2 |
| Armenia | 1 | 1 | 2 | 2 | 7 | 1 | 1 | 2 | 2 |
| Australia | 3 | 2 | 3 | 3 | 12 | 2 | 1 | 3 | 3 |
| Austria | 3 | 2 | 3 | 3 | 1 | 2 | 1 | 3 | 3 |
| Azerbaijan | 1 | 1 | 2 | 2 | 2 | 1 | 1 | 2 | 2 |
| Bahrain | 1 | 1 | 2 | 2 | 9 | 1 | 1 | 2 | 2 |
| Bangladesh | 2 | 3 | 1 | 1 | 8 | 3 | 1 | 1 | 1 |
| Belarus | 1 | 1 | 2 | 2 | 6 | 1 | 1 | 2 | 2 |
| Belgium | 3 | 2 | 3 | 3 | 1 | 2 | 1 | 3 | 3 |
| Benin | 2 | 3 | 1 | 1 | 4 | 3 | 1 | 1 | 1 |
| Bhutan | 1 | 1 | 2 | 2 | 2 | 1 | 1 | 2 | 2 |
| Bolivia | 1 | 1 | 2 | 2 | 11 | 1 | 1 | 2 | 2 |
| Bosnia and Herzegovina | 1 | 1 | 2 | 2 | 7 | 1 | 1 | 2 | 2 |
| Botswana | 1 | 1 | 2 | 2 | 2 | 1 | 1 | 2 | 2 |
| Brazil | 1 | 1 | 2 | 2 | 9 | 1 | 1 | 2 | 2 |
| Bulgaria | 1 | 1 | 2 | 2 | 6 | 1 | 1 | 2 | 2 |
| Burkina Faso | 2 | 3 | 1 | 1 | 13 | 3 | 1 | 1 | 1 |
| Burundi | 2 | 3 | 1 | 1 | 5 | 3 | 1 | 1 | 1 |
| Cambodia | 1 | 1 | 2 | 2 | 11 | 1 | 1 | 2 | 1 |
| Cameroon | 2 | 3 | 1 | 1 | 4 | 3 | 1 | 1 | 1 |
| Canada | 3 | 2 | 3 | 3 | 12 | 2 | 1 | 3 | 3 |
| Central African Republic | 2 | 3 | 1 | 1 | 3 | 3 | 0 | 1 | 1 |
| Chad | 2 | 3 | 1 | 1 | 4 | 3 | 1 | 1 | 1 |
| Chile | 1 | 1 | 2 | 2 | 9 | 1 | 1 | 2 | 2 |
| China | 1 | 1 | 2 | 2 | 11 | 1 | 1 | 2 | 2 |
| Colombia | 1 | 1 | 2 | 2 | 11 | 1 | 1 | 2 | 2 |
| Comoros | 2 | 3 | 1 | 1 | 5 | 3 | 1 | 1 | 1 |
| Congo (Brazzaville) | 2 | 3 | 1 | 1 | 4 | 3 | 1 | 1 | 1 |

**Appendix** (continued) Cluster membership for each algorithm

| Country | *k*-means | *k*-mean++ | *k*-medoids (PAM) | CLARA | Affinity propagation | Spectral clustering | DBSCAN | AGNES | DIANA |
|---|---|---|---|---|---|---|---|---|---|
| Congo (Kinshasa) | 2 | 3 | 1 | 1 | 13 | 3 | 1 | 1 | 1 |
| Costa Rica | 1 | 1 | 2 | 2 | 9 | 1 | 1 | 2 | 2 |
| Croatia | 1 | 1 | 2 | 2 | 7 | 1 | 1 | 2 | 2 |
| Cyprus | 1 | 1 | 2 | 2 | 9 | 1 | 1 | 2 | 2 |
| Czech Republic | 1 | 1 | 2 | 2 | 9 | 1 | 1 | 2 | 2 |
| Denmark | 3 | 2 | 3 | 3 | 12 | 2 | 1 | 3 | 3 |
| Dominican Republic | 1 | 1 | 2 | 2 | 11 | 1 | 1 | 2 | 2 |
| Ecuador | 1 | 1 | 2 | 2 | 11 | 1 | 1 | 2 | 2 |
| Egypt | 1 | 1 | 2 | 2 | 7 | 3 | 1 | 2 | 1 |
| El Salvador | 1 | 1 | 2 | 2 | 11 | 1 | 1 | 2 | 2 |
| Estonia | 3 | 2 | 2 | 2 | 1 | 1 | 1 | 3 | 3 |
| Ethiopia | 2 | 3 | 1 | 1 | 13 | 3 | 1 | 1 | 1 |
| Finland | 3 | 2 | 3 | 3 | 12 | 2 | 1 | 3 | 3 |
| France | 3 | 2 | 3 | 3 | 1 | 2 | 1 | 3 | 3 |
| Gabon | 1 | 1 | 2 | 2 | 7 | 1 | 1 | 2 | 2 |
| Gambia | 2 | 3 | 1 | 1 | 4 | 3 | 1 | 1 | 1 |
| Georgia | 2 | 3 | 1 | 1 | 8 | 3 | 1 | 1 | 1 |
| Germany | 3 | 2 | 3 | 3 | 1 | 2 | 1 | 3 | 3 |
| Ghana | 2 | 3 | 1 | 1 | 13 | 3 | 1 | 1 | 1 |
| Greece | 1 | 1 | 2 | 2 | 7 | 1 | 1 | 2 | 2 |
| Guatemala | 1 | 1 | 2 | 2 | 11 | 1 | 1 | 2 | 2 |
| Guinea | 2 | 3 | 1 | 1 | 4 | 3 | 1 | 1 | 1 |
| Haiti | 2 | 3 | 1 | 1 | 5 | 3 | 1 | 1 | 1 |
| Honduras | 1 | 1 | 2 | 2 | 11 | 1 | 1 | 2 | 2 |
| Hong Kong | 3 | 2 | 3 | 3 | 1 | 2 | 1 | 3 | 3 |
| Hungary | 1 | 1 | 2 | 2 | 6 | 1 | 1 | 2 | 2 |
| Iceland | 3 | 2 | 3 | 3 | 1 | 2 | 1 | 3 | 3 |
| India | 2 | 3 | 1 | 1 | 8 | 3 | 1 | 1 | 1 |
| Indonesia | 1 | 1 | 2 | 2 | 11 | 1 | 1 | 2 | 2 |
| Iran | 1 | 1 | 2 | 2 | 8 | 1 | 1 | 1 | 2 |
| Iraq | 2 | 3 | 1 | 1 | 7 | 3 | 1 | 2 | 1 |
| Ireland | 3 | 2 | 3 | 3 | 12 | 2 | 1 | 3 | 3 |
| Israel | 1 | 1 | 2 | 2 | 9 | 1 | 1 | 2 | 2 |
| Italy | 1 | 1 | 2 | 2 | 6 | 1 | 1 | 2 | 2 |
| Ivory Coast | 2 | 3 | 1 | 1 | 4 | 3 | 1 | 1 | 1 |
| Jamaica | 1 | 1 | 2 | 2 | 11 | 1 | 1 | 2 | 2 |
| Japan | 1 | 1 | 2 | 2 | 1 | 1 | 1 | 3 | 3 |
| Jordan | 1 | 1 | 2 | 2 | 11 | 1 | 1 | 2 | 2 |
| Kazakhstan | 1 | 1 | 2 | 2 | 9 | 1 | 1 | 2 | 2 |
| Kenya | 2 | 3 | 1 | 1 | 13 | 3 | 1 | 1 | 1 |
| Kosovo | 1 | 1 | 2 | 2 | 11 | 1 | 1 | 2 | 2 |
| Kuwait | 1 | 1 | 2 | 2 | 9 | 1 | 1 | 2 | 2 |
| Kyrgyzstan | 1 | 1 | 2 | 2 | 11 | 1 | 1 | 2 | 2 |
| Laos | 2 | 3 | 1 | 1 | 2 | 3 | 1 | 1 | 1 |
| Latvia | 1 | 1 | 2 | 2 | 6 | 1 | 1 | 2 | 2 |
| Lebanon | 1 | 1 | 2 | 2 | 7 | 1 | 1 | 2 | 2 |
| Lesotho | 2 | 3 | 1 | 1 | 13 | 3 | 1 | 1 | 1 |
| Liberia | 2 | 3 | 1 | 1 | 13 | 3 | 1 | 1 | 1 |
| Libya | 1 | 1 | 2 | 2 | 2 | 1 | 1 | 2 | 2 |
| Lithuania | 1 | 1 | 2 | 2 | 6 | 1 | 1 | 2 | 2 |
| Luxembourg | 3 | 2 | 3 | 3 | 12 | 2 | 1 | 3 | 3 |
| Madagascar | 2 | 3 | 1 | 1 | 5 | 3 | 1 | 1 | 1 |
| Malawi | 2 | 3 | 1 | 1 | 4 | 3 | 1 | 1 | 1 |
| Malaysia | 1 | 1 | 2 | 2 | 9 | 1 | 1 | 2 | 2 |
| Mali | 2 | 3 | 1 | 1 | 13 | 3 | 1 | 1 | 1 |
| Malta | 3 | 2 | 3 | 3 | 1 | 2 | 1 | 3 | 3 |
| Mauritania | 2 | 3 | 1 | 1 | 5 | 3 | 1 | 1 | 1 |
| Mauritius | 1 | 1 | 2 | 2 | 9 | 1 | 1 | 2 | 2 |
| Mexico | 1 | 1 | 2 | 2 | 9 | 1 | 1 | 2 | 2 |
| Moldova | 1 | 1 | 2 | 2 | 7 | 1 | 1 | 2 | 2 |
| Mongolia | 1 | 1 | 2 | 2 | 6 | 1 | 1 | 2 | 2 |
| Montenegro | 1 | 1 | 2 | 2 | 6 | 1 | 1 | 2 | 2 |
| Morocco | 2 | 3 | 2 | 2 | 8 | 3 | 1 | 1 | 1 |
| Mozambique | 2 | 3 | 1 | 1 | 4 | 3 | 1 | 1 | 1 |
| Myanmar | 1 | 1 | 2 | 2 | 2 | 3 | 1 | 1 | 1 |
| Namibia | 1 | 1 | 2 | 2 | 2 | 1 | 1 | 2 | 1 |
| Nepal | 1 | 1 | 2 | 2 | 2 | 3 | 1 | 2 | 1 |
| Netherlands | 3 | 2 | 3 | 3 | 12 | 2 | 1 | 3 | 3 |
| New Zealand | 3 | 2 | 3 | 3 | 12 | 2 | 1 | 3 | 3 |
| Nicaragua | 1 | 1 | 2 | 2 | 2 | 1 | 1 | 2 | 2 |
| Niger | 2 | 3 | 1 | 1 | 4 | 3 | 1 | 1 | 1 |
| Nigeria | 2 | 3 | 1 | 1 | 13 | 3 | 1 | 1 | 1 |
| North Macedonia | 1 | 1 | 2 | 2 | 7 | 1 | 1 | 2 | 2 |
| Northern Cyprus | 1 | 1 | 2 | 2 | 1 | 1 | 1 | 3 | 2 |

**Appendix** (continued) Cluster membership for each algorithm

| Country | *k*-means | *k*-mean++ | *k*-medoids (PAM) | CLARA | Affinity propagation | Spectral clustering | DBSCAN | AGNES | DIANA |
|---|---|---|---|---|---|---|---|---|---|
| Norway | 3 | 2 | 3 | 3 | 12 | 2 | 1 | 3 | 3 |
| Pakistan | 2 | 3 | 1 | 1 | 4 | 3 | 1 | 1 | 1 |
| Palestinian Territories | 1 | 1 | 2 | 2 | 7 | 3 | 1 | 2 | 1 |
| Panama | 1 | 1 | 2 | 2 | 9 | 1 | 1 | 2 | 2 |
| Paraguay | 1 | 1 | 2 | 2 | 11 | 1 | 1 | 2 | 2 |
| Peru | 1 | 1 | 2 | 2 | 11 | 1 | 1 | 2 | 2 |
| Philippines | 1 | 1 | 2 | 2 | 2 | 1 | 1 | 2 | 2 |
| Poland | 1 | 1 | 2 | 2 | 9 | 1 | 1 | 2 | 2 |
| Portugal | 1 | 1 | 2 | 2 | 9 | 1 | 1 | 2 | 2 |
| Qatar | 1 | 1 | 2 | 2 | 9 | 1 | 1 | 3 | 2 |
| Romania | 1 | 1 | 2 | 2 | 11 | 1 | 1 | 2 | 2 |
| Russia | 1 | 1 | 2 | 2 | 6 | 1 | 1 | 2 | 2 |
| Rwanda | 2 | 3 | 1 | 1 | 10 | 3 | 0 | 1 | 1 |
| Saudi Arabia | 1 | 1 | 2 | 2 | 9 | 1 | 1 | 2 | 2 |
| Senegal | 2 | 3 | 1 | 1 | 13 | 3 | 1 | 1 | 1 |
| Serbia | 1 | 1 | 2 | 2 | 6 | 1 | 1 | 2 | 2 |
| Sierra Leone | 2 | 3 | 1 | 1 | 4 | 3 | 1 | 1 | 1 |
| Singapore | 3 | 2 | 3 | 3 | 12 | 2 | 1 | 3 | 3 |
| Slovakia | 1 | 1 | 2 | 2 | 6 | 1 | 1 | 2 | 2 |
| Slovenia | 1 | 1 | 2 | 2 | 9 | 1 | 1 | 2 | 2 |
| Somalia | 2 | 3 | 1 | 1 | 10 | 3 | 0 | 1 | 1 |
| South Africa | 1 | 1 | 2 | 2 | 2 | 1 | 1 | 2 | 2 |
| South Korea | 1 | 1 | 2 | 2 | 7 | 1 | 1 | 2 | 2 |
| South Sudan | 2 | 3 | 1 | 1 | 5 | 3 | 1 | 1 | 1 |
| Spain | 1 | 1 | 2 | 2 | 9 | 1 | 1 | 2 | 2 |
| Sri Lanka | 1 | 1 | 2 | 2 | 11 | 1 | 1 | 2 | 2 |
| Swaziland | 2 | 3 | 1 | 1 | 4 | 3 | 1 | 1 | 1 |
| Sweden | 3 | 2 | 3 | 3 | 12 | 2 | 1 | 3 | 3 |
| Switzerland | 3 | 2 | 3 | 3 | 12 | 2 | 1 | 3 | 3 |
| Syria | 2 | 3 | 1 | 1 | 5 | 3 | 1 | 1 | 1 |
| Taiwan | 1 | 1 | 2 | 2 | 6 | 1 | 1 | 2 | 2 |
| Tajikistan | 1 | 1 | 2 | 2 | 2 | 3 | 1 | 2 | 1 |
| Tanzania | 2 | 3 | 1 | 1 | 4 | 3 | 1 | 1 | 1 |
| Thailand | 1 | 1 | 2 | 2 | 11 | 1 | 1 | 2 | 2 |
| Togo | 2 | 3 | 1 | 1 | 4 | 3 | 1 | 1 | 1 |
| Trinidad & Tobago | 1 | 1 | 2 | 2 | 9 | 1 | 1 | 2 | 2 |
| Tunisia | 1 | 1 | 2 | 2 | 7 | 1 | 1 | 2 | 2 |
| Turkey | 1 | 1 | 2 | 2 | 6 | 1 | 1 | 2 | 2 |
| Turkmenistan | 1 | 1 | 2 | 2 | 2 | 1 | 1 | 2 | 2 |
| Uganda | 2 | 3 | 1 | 1 | 13 | 3 | 1 | 1 | 1 |
| Ukraine | 1 | 1 | 2 | 2 | 7 | 1 | 1 | 2 | 2 |
| United Arab Emirates | 3 | 2 | 2 | 2 | 1 | 1 | 1 | 3 | 3 |
| United Kingdom | 3 | 2 | 3 | 3 | 1 | 2 | 1 | 3 | 3 |
| United States | 1 | 1 | 2 | 2 | 9 | 1 | 1 | 3 | 2 |
| Uruguay | 1 | 1 | 2 | 2 | 1 | 1 | 1 | 3 | 3 |
| Uzbekistan | 3 | 2 | 3 | 3 | 2 | 2 | 0 | 2 | 3 |
| Venezuela | 1 | 1 | 2 | 2 | 7 | 1 | 1 | 2 | 2 |
| Vietnam | 1 | 1 | 2 | 2 | 11 | 1 | 1 | 2 | 2 |
| Yemen | 2 | 3 | 1 | 1 | 13 | 3 | 1 | 1 | 1 |
| Zambia | 2 | 3 | 1 | 1 | 13 | 3 | 1 | 1 | 1 |
| Zimbabwe | 2 | 3 | 1 | 1 | 13 | 3 | 1 | 1 | 1 |