Scientific Research Publishing

Open Journal of Applied Sciences, 2017, 7, 661-690
http://www.scirp.org/journal/ojapps
ISSN Online: 2165-3925
ISSN Print: 2165-3917

The Analysis of Human Development Index (HDI) for Categorizing the Member States of the United Nations (UN)

Sivarajah Mylevaganam

Alumnus, Spatial Sciences Laboratory, Texas A&M University, College Station, USA
Email: sivaloga@hushmail.com

How to cite this paper: Mylevaganam, S. (2017) The Analysis of Human Development Index (HDI) for Categorizing the Member States of the United Nations (UN). *Open Journal of Applied Sciences*, **7**, 661-690.
https://doi.org/10.4236/ojapps.2017.712048

Received: November 10, 2017
Accepted: December 17, 2017
Published: December 21, 2017

Copyright © 2017 by author and Scientific Research Publishing Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/

Open Access

Abstract

To categorize the nations to reflect the development status, to date, there are many conceptual frameworks. The Human Development index (HDI) that is published by the United Nations Development Programme is widely accepted and practiced by many people such as academicians, politicians, and donor organizations. However, though the development of HDI has gone through many revisions since its formulation in 1990, even the current version of the index formulation published in 2016 needs research to better understand and to gap-fill the knowledge base that can enhance the index formulation to facilitate the direction of attention such as release of funds. Therefore, in this paper, based on principal component analysis and K-means clustering algorithm, the data that reflect the measures of life expectancy index (LEI), education index (EI), and income index (II) are analyzed to categorize and to rank the member states of the UN using R statistical software package, an open source extensible programming language for statistical computing and graphics. The outcome of the study shows that the proportion of total eigen value (*i.e.*, proportion of total variance) explained by PCA-1 (*i.e.*, first principal component) accounts for more than 85% of the total variation. Moreover, the proportion of total eigen value explained by PCA-1 increases with time (*i.e.*, yearly) though the amount of increase with time is not significant. However, the proportions of total eigen value explained by PCA-2 and PCA-3 decrease with time. Therefore, the loss of information in choosing PCA-1 to represent the chosen explanatory variables (*i.e.*, LEI, EI, and II) may diminish with time if the trend of increasing pattern of proportion of total eigen value explained by PCA-1 with time continues in the future as well. On the other hand, the correlation between EI and PCA-1 increases with time although the magnitude of increase is not that significant. This same trend is observed in II as well. However, in contrast to these observations, the correlation between PCA-1 and LEI decreases with time. These findings imply that the contribu-

tions of EI and II to PCA-1 increase with time, but the contribution of LEI to PCA-1 decreases with time. On top of these, as per Hopkins statistic, the clusterability of the information conveyed by PCA-1 alone is far better than the clusterability of the information conveyed by PCA scores (*i.e.*, PCA-1, PCA-2, and PCA-3) and the explanatory variables. Therefore, choosing PCA-1 to represent the chosen explanatory variables is becoming more concrete.

Keywords

Human Development Index, Economy, Sustainability, United Nations Development Programme, Education, Life Expectancy, Per Capita Income, JavaScript, R Statistical Software, Principal Component Analysis, K-Means Clustering, Hopkins Statistic

1. Introduction

Since the incipient of relativity theory, the categorization of parameters of interest has been surfaced in many fields, including in the scientific field. This has forced many apex bodies such as donor organizations (e.g., World Bank and United Nations Development Programme) to find some schemes to pool the countries into few categories to summarize the status of development, which can help to facilitate the direction of attention such as release of funds. Among all the categorization schemes practiced, the Human Development Index (HDI) [1] [2] [3] that is published annually by the human development office of the United Nations Development Programme (UNDP) and geared towards people centered policies has become one of the well accepted measures to categorize the member states of the United Nations (UN) into few tiers [1] [2] [3].

The HDI, which is to evaluate the development of a UN country from the perspective of well-being of human-beings, in addition to the economic advancement, is basically an index composed of three measures, namely life expectancy, education, and per capita income [1] [2]. These three measures are defined through three indices: Life expectancy index (LEI), education index (EI), and income index (II) [1] [2]. The LEI is used to measure the population health and the longevity. The EI is a measure of education and the access to knowledge. On the other hand, the II is to measure the standard of living. Though HDI has been widely used as the measure of development [1]-[8], many issues in the fundamental formulation of the underlying concept have been researched [3]-[8].

Stanton (2007) reviews the key issues on HDI into five categories: Poor data, incorrect choice of indictors, incorrect specification of income, redundancy, and formulation of HDI. [3] [4] [5] point out the quality of data and the frequency of data collection. [6] identifies three sources of data error which are due to data updating, formula revisions, and thresholds to classify a country's development status, to formulate and propose a statistical framework to calculate country spe-

cific measures of data uncertainty and its impact on rank assignments. Based on principle component analysis, [7] [8] provide a theoretical support for the HDI ranking system. The inclusion and the omission of the components in the index formulation has been criticized by [3].

To address some of the issues raised, with improved quality of data, the development of HDI has gone through many revisions since its formulation in 1990 [3]. However, even the current version of the index formulation published in 2016 okneeds research to better understand and to gap-fill the knowledge base that can enhance the index formulation to facilitate the direction of attention and for simplifying problems. As per the current version of the index, the HDI index is basically a multiple linear regression equation. The explanatory variables in the equations are the logarithms of LEI, EI, and II. However, the weights that determine the strength of the explanatory variables are given equal values (*i.e.*, 1/3). In other words, it is assumed that the index is defined by setting equal values to the measures chosen: LEI, EI, and II. Therefore, with the current version of the index formulation, country-A (LEI = 0.5, EI = 0.5, II = 1.0) and country-B (LEI = 1.0, EI = 0.5, II = 0.5) will have the same HDI index value (*i.e.*, $HDI_{country\text{-}A} = (0.5 * 0.5 * 1.0)^{1/3} = HDI_{country\text{-}B} = (1.0 * 0.5 * 0.5)^{1/3} = 0.62996$) though they may be heading in two different directions as depicted by the values of LEI, EI, and II. Moreover, the current literature does not critically analyze the components that form the HDI with time, and the critical or the cut-off HDI values that are used to categorize the nations into tiers, to gap-fill the knowledge base that can enhance the index formulation in line with UNDP that the development of HDI should be seen as evolving and improving with active participation of its users, rather than as something cast in stone [3].

Having said this, the information from the data that represent the measures of LEI, EI, and II could be potentially maximized with the help of some of the well-established clustering algorithms and multivariate analysis techniques [9] [10] [11]. In fact, this will avoid the use of any form of ad-hoc combinations such as geometric or linear combination. Instead, the countries will be clustered based on the underlying data chosen to reflect the people centered policies. Moreover, the well-established clustering algorithms and multivariate analysis techniques may also lead to categorize the countries within a region or cluster of interest into few tiers solely based on those countries' LEI, EI, and II. This will, in turn, help to categorize the member states of the UN from local and global perspective. Therefore, the objective of this paper is to analyze the data that reflect the measures of life expectancy index, education index, and income index based on principal component analysis and K-means clustering algorithm to understand the trends of the indices that form HDI with time; to categorize; and to rank the member states of the UN using R statistical software package, an open source extensible programming language for statistical computing and graphics.

2. Human Development Index

The HDI reported by the human development office of the UNDP is basically an

index composed of three measures, namely life expectancy, education, and per capita income [1] [2]. These measures are defined through three indices: LEI, EI, and II [1] [2]. The underlying theoretical formulation of HDI developed by a few economists is shown in **Figure 1** [1] [2].

2.1. Life Expectancy Index

The LEI that measures the population health and the longevity is defined by:

$$\text{LEI} = \frac{\text{LE} - 20}{85 - 20} \tag{1}$$

where LE is the life expectancy at birth in years. The value of LEI varies between 0 and 1. The LEI becomes 1 if LE is 85. Similarly, the LEI becomes 0 if LE is 20.

2.2. Education Index

The EI measures the level of education and the access to knowledge. As shown in Equation (2), the EI is formed of two indices, namely mean years of schooling index (MYSI) and expected years of schooling index (EYSI).

$$\text{EI} = \frac{\text{MYSI} + \text{EYSI}}{2} = 0.5 * \frac{\text{MYS}}{15} + 0.5 * \frac{\text{EYS}}{18} \tag{2}$$

where MYS and EYS are the mean years of schooling (years) and the expected years of schooling (years), respectively. The value of EI varies between 0 and 1.

2.3. Income Index

The II that measures the standard of living is defined by:

$$\text{II} = \frac{\ln(GDP) - \ln(100)}{\ln(75000) - \ln(100)} \tag{3}$$

where GDP is the gross domestic product per capita (2011 PPP $). The value of II varies between 0 and 1. The II becomes 1 if GDP is 75000. Similarly, the LEI becomes 0 if LE is 100.

2.4. Computation of Human Development Index

The theoretical formulation of HDI is given by:

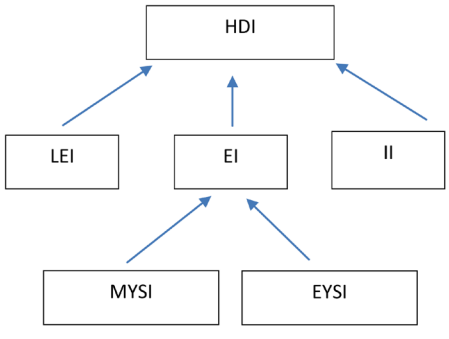


Figure 1. The conceptual framework of HDI.

$$\mathrm{HDI} = \left(\mathrm{LEI} * \mathrm{EI} * \mathrm{II}\right)^{1/3} \tag{4}$$

Taking the logarithm of Equation (4):

$$\ln \mathrm{HDI} = \frac{1}{3}\left(\ln \mathrm{LEI} + \ln \mathrm{EI} + \ln \mathrm{II}\right) \tag{4'}$$

With further simplification,

$$Y = a_1 X_1 + a_2 X_2 + a_3 X_3 \tag{4"}$$

where $a_1 = a_2 = a_3 = \frac{1}{3}$ and $X_1 = \ln \mathrm{LEI}$, $X_2 = \ln \mathrm{EI}$ and $X_3 = \ln \mathrm{II}$. Therefore, the Equation (4-4") is basically a multiple linear regression equation. The explanatory variables in the equations are the logarithms of LEI, EI, and II. However, the weights that determine the strengths of the explanatory variables are given equal values.

To demonstrate the computation of HDI, the values reported for the measures discussed in Sections 2.1-2.3 are extracted for Norway, one of the UN member states. These values are extracted from the human development report published in 2016 [1]. The extracted values are placed in **Table 1**.

As per the current methodology implemented by UNDP, the values of LEI, II, and EI are $\frac{81.7-20}{85-20} = 0.949$, $\frac{\ln(67614)-\ln(100)}{\ln(75000)-\ln(100)} = 0.984$, $0.5*\frac{12.7}{15} + 0.5*\frac{17.7}{18} = 0.915$, respectively. Therefore, the value of HDI for this country is $\left(\mathrm{LEI} * \mathrm{EI} * \mathrm{II}\right)^{1/3} = 0.949$.

3. Methodology

In this section of the manuscript, the methodologies adopted to analyze the data that reflect the measures of life expectancy index, education index, and income index; to categorize; and to rank the member states of the UN are presented. The Section 3.1 outlines the fundamental behind the principal component analysis which is used to reduce the complexity of multidimensional data; the Section 3.2 outlines the clustering algorithm, particularly the k-means clustering algorithm that is used to form groups or clusters such that the variation within each cluster is minimized; the Section 3.3 outlines the development of virtual human development index; and the Section 3.4 presents the implementation of virtual human development index using R statistical programming language.

Table 1. The measures of HDI in Norway in 2015.

Measures of HDI	Reported Values in 2015
Life expectancy at birth (years)	81.7
Mean years of schooling (years)	12.7
Expected years of schooling (years)	17.7
Gross domestic product (GDP) per capita (2011 PPP $)	67614

3.1. Principal Component Analysis

Principal component analysis (PCA) is one of the techniques or tools used in information theory to reduce dimensions in a multidimensional data. In other words, PCA is used to reduce the complexity of multidimensional data to enhance the handling, analysis, interpretation, and visualization of multidimensional data [9] [10] [11]. In simple terms, PCA helps to define the multidimensional data using principal component axes, instead of conventional coordinate system such as Cartesian coordinate system (*i.e.*, x − y − z). However, the principal component axes are formed in way that the variance in the multidimensional data is maximized along the principal component axes.

To illustrate the concept of PCA, for example, consider the data points shown in **Figure 2(a)**. The data points are shown in a 2-D Cartesian coordinate system. On the other hand, in **Figure 2(b)**, the given data points are shown using principal component axes. As can be observed, basically, PCA transforms the multidimensional to a new coordinate system. However, the coordinate system is formed in a way that the variance in the multidimensional data is maximized along the principal component axes. Moreover, the number of principal component axes is equal to the dimension of the multidimensional data. In the considered problem, there are two variables (*i.e.*, X and Y). Therefore, there are two principal component axes as shown in **Figure 2(b)**.

3.1.1. Mathematical Formulation of PCA

Consider the below shown multidimensional data vector (X) with three variables:

$$X = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}$$

The principal component scores (*i.e.*, multidimensional data in the new coordinate system) for this multidimensional data are given by Equations (5)-(7).

$$Y_1 = e_{11}X_1 + e_{12}X_2 + e_{13}X_3 \tag{5}$$

$$Y_2 = e_{21}X_1 + e_{22}X_2 + e_{23}X_3 \tag{6}$$

$$Y_3 = e_{31}X_1 + e_{32}X_2 + e_{33}X_3 \tag{7}$$

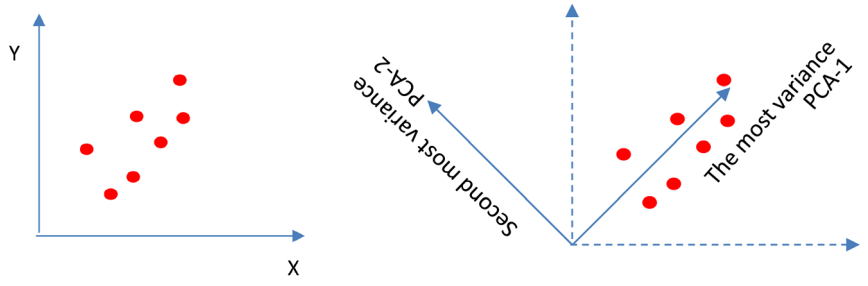


Figure 2. Formation of principal component axes.

As can be observed, the principal component scores (*i.e.* Y_1, Y_2 and Y_3) are the linear combinations of the given multidimensional data without an intercept. The coefficients e_{11}, e_{12}, e_{13}, e_{21}, e_{22}, e_{23}, e_{31}, e_{32}, and e_{33} are the regression coefficients. Since there are three variables (*i.e.* X_1, X_2 and X_3) involved in this problem, there should be three principal components. The first principal component is obtained by maximizing the variance of Y_1 subjected that the sum of squared coefficients is equal to one (*i.e.*, $e_{11}^2 + e_{12}^2 + e_{13}^2 = 1$). However, when finding the remaining principal components, in addition to maximizing the variance and setting the sum of squared coefficients to one, it is also ensured that the correlation between the principal component of interest (e.g., second principal component) and previous principal components (e.g., first principal component) is zero. The regression coefficients and the variance of the principal components are found using the eigen vectors (x) and eigen values (λs) of the variance-covariance matrix (A, *i.e.*, variance-covariance matrix of the given multidimensional data), respectively, as shown in Equations (8)-(9). The solution of Equation (9) leads to solve the Equation (8). Moreover, it is noted that the number of eigen vectors is equal to the dimension of the multidimensional data, and every eigen vector is associated with an eigen value.

$$Ax = \lambda x \tag{8}$$

$$\left|A - \lambda I\right| = 0 \tag{9}$$

where "I" is the identity matrix.

3.1.2. Reduction of Dimensions Using PCA

As explained in Section 3.1.1, each principal component is associated with an eigen value. Moreover, the eigen value (λ) is the amount of variance associated with the principal component. Therefore, the proportion of total variation explained by a given principal component (say i) and the cumulative proportion of total variation explained by a given principal component are given by Equation (10) and Equation (11), respectively.

Proportion of Total Variation of Principal Component "i"

$$= \frac{\lambda_i}{\lambda_1 + \lambda_2 + \cdots + \lambda_p} \tag{10}$$

Cumulative Proportion of Total Variation of Principal Component "i"

$$= \frac{\lambda_1 + \cdots + \lambda_i}{\lambda_1 + \lambda_2 + \cdots + \lambda_p} \tag{11}$$

where λ_i is the amount of variance associated with the principal component "i" and "p" is the number of principal components (*i.e.*, dimension of the data). The proportion of total variation explained by a given principal component and the cumulative proportion of total variation explained by a given principal component are used to determine the number of dimensions required to express the problem of interest, considering the correlations between the variables in the multidimensional data.

3.1.3. Computation of PCA for a 2-Dimensional Data

To demonstrate the computation of PCA for a 2-Dimensional data, consider the multidimensional data of two variables that are extracted from the report published in 2016 [1] and placed in Table 2. The variables X and Y represent the education index and the life expectancy index, respectively.

To compute the PCA scores, at first, the variance-covariance matrix is computed. For the given multidimensional data, the variance-covariance matrix is given by

	X	Y
X	0.02500196	0.02057909
Y	0.02057909	0.02224321

The eigen vectors and the eigenvalues associated with the variance-covariance matrix is obtained following Equations (8)-(9). For the given multidimensional data, the eigen vectors are given by

Eigen Vector-1	Eigen Vector-2
+0.7303690	+0.6830528
+0.6830528	−0.7303690

The sum of squared coefficients for the first principal component (PCA-1) is equal to $0.7303690^2 + 0.6830528^2 = 1$. Similarly, the sum of squared coefficients for the second principal component (PCA-2) is equal to $0.6830528^2 + (-0.7303690)^2 = 1$. The eigen values associated with eigen vector-1 and eigen vector-2 are 0.044247849 and 0.002997318, respectively. Therefore, as per the computed eigen values, the maximum variation is explained by PCA-1. The proportion of variation explained by PCAs are placed in Table 3. Since the variation explained by the first principal component is very high (0.9365582)

Table 2. Multidimensional data of two variables extracted from the report published in 2016 [1].

	X	Y		X	Y	Summary
1	0.398	0.626	6	0.694	0.865	
2	0.715	0.892	7	0.808	0.869	Mean (X) = 0.6962
3	0.658	0.847	8	0.73	0.844	Mean (Y) = 0.8301 Var (X) = 0.025
4	0.718	0.946	9	0.939	0.962	Var (Y) = 0.022
5	0.482	0.503	10	0.82	0.947	

Table 3. Proportion and cumulative proportion of total variance of principal components.

Principal Component	Eigen Value (Variance)	Proportion of Total Variance	Cumulative Proportion of Total Variance
1	0.044247849	0.9365582	0.9365582
2	0.002997318	0.06344179	1.0000000

compared to the second principal component (0.06344179), without loss of significant information, the given multidimensional dataset could be reduced to one dimension that is explained by PCA-1.

The principal component scores are obtained by following Equations (5)-(7). In PCA, it is a widespread practice to use the difference between the variables and their sample means, instead of using the raw data. Therefore, considering the first row (*i.e.*, $X = 0.398$ and $Y = 0.626$) in the given data, the mean adjusted first principal component score is given by

$$Y_1 = +0.7303690 * X_1 + 0.6830528 * X_2$$

$$Y_1 = +0.7303690 * (0.398 - 0.6962) + 0.6830528 * (0.626 - 0.8301)$$
$$= -0.35720711$$

Similarly, the second principal component is given by

$$Y_2 = +0.6830528 * X_1 - 0.7303690 * X_2$$

$$Y_2 = +0.6830528 * (0.398 - 0.6962) - 0.7303690 * (0.626 - 0.8301)$$
$$= -0.0546180239$$

Following the same procedure, the new dataset for the given multidimensional data is placed in Table 4.

Since the variables in multidimensional data may not have the same units, to enhance the interpretation of PCA, often, the multidimensional data is standardized.

3.2. Measure of Clusterability Using Hopkins Statistic

Hopkins statistic which is a statistical hypothetical test measures the clusterability (*i.e.*, cluster tendency) of a given dataset [12] [13]. The null hypothesis of Hopkins statistic checks if the given dataset comes from a uniform distribution. To test the null hypothesis, at first, few points are uniformly selected from the given dataset. The distances between these points and their closest nearest points are computed ($d_{\text{actual},i}$). Similarly, few points are uniformly selected from a random dataset. The distances between these points and their closest nearest points in the actual dataset are computed ($d_{\text{random},i}$) [12] [13]. Then, the Hopkins statistic is found using the following equation:

$$\text{Hopkins Statistic} = \frac{\sum d_{\text{random},i}}{\sum d_{\text{actual},i} + \sum d_{\text{random},i}} \tag{12}$$

Table 4. Principal component scores of the multidimensional data.

	PCA Score-1	PCA Score-2		PCA Score-1	PCA Score-2
1	−0.35720711	−0.0546180239	6	0.02223173	−0.0269925947
2	0.05601190	−0.0323684497	7	0.10822601	0.0479539464
3	−0.01635650	−0.0384358526	8	0.03418091	0.0129350547
4	0.09508786	−0.0697592181	9	0.26742826	0.0695095426
5	−0.37987161	0.0925937985	10	0.17026855	−0.0008182033

where $d_{\text{actual},i}$ is the distance between the "i"th point and its closest nearest point. If the given dataset comes from a uniform distribution, Hopkins statistic will be towards zero. Therefore, a dataset with a Hopkins statistic that is greater than zero (ideally above 0.5) is considered clusterable.

3.3. K-Means Clustering Algorithm

Clustering is one of the techniques or information tools used to group a set of multidimensional data. In other words, using clustering algorithms, the multidimensional data is pooled into a set of groups such that the data within a group behaves in an equivalent or homogeneous manner, but the data between groups behave dissimilarly [9] [11].

K-means clustering algorithm, which is an unsupervised algorithm, is the simplest and most widely used clustering algorithm to partition n observations into $k\,(<n)$ clusters. In other words, given a set of observations $(y_1, y_2, y_3, \cdots, y_n)$, K-means algorithm partitions the observations into k clusters $(G_1, G_2, G_3, \cdots, G_k)$ such that the total variation within each clusters is kept at minimum. This is accomplished by minimizing the sum of squared distances (e.g., Euclidean distance, Manhattan distance, and Correlation based distance) from the data points within a cluster to the centroid of the cluster [9] [11].

To illustrate the concept of K-means clustering algorithm, consider the data points shown in **Figure 3**. These data points are extracted from [2]. If the objective is to cluster these points into two clusters, at first, two of the given data points are considered as the centroids of the clusters. If the datasets for country-1 and country-2 are assumed to be the initial guesses, the centroids of the clusters are [LEI = 0.948, EI = 0.916, II = 0.983] and [LEI = 0.858, EI = 0.631, II = 0.730]. These could be visualized as points lying on a 3-D space whose X, Y, and Z coordinates are denoted by LEI, EI, and II, respectively. The remaining data points are assigned to one of these clusters such that the distance (it is emphasized that the term distance does not refer to the spatial distance corresponding to the geographical coordinates of the data points) between the data point and the centroid of the cluster is minimum. For example, as shown in **Figure 3**, if $d_1 < d_2$, country-3 will be assigned to the cluster that is centered at

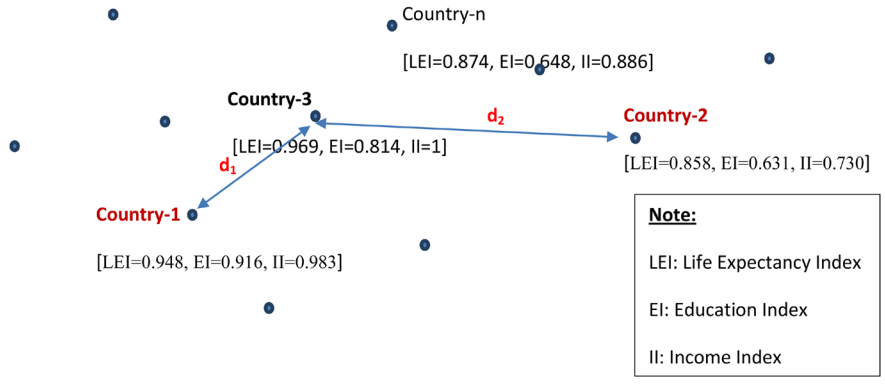


Figure 3. The implementation of k-mean clustering algorithm using LEI, EI, and II.

country-1. Otherwise, country-3 belongs to country-2. Having assigned the data points to one of the two clusters, the centroids of the clusters are updated. This procedure is repeated until the centroids of the clusters do not change with further iteration.

3.3.1. Optimum Number of Clusters

To determine the optimum number of clusters, as underscored in the literature, there are many methods available such as elbow, silhouette, and gap statistic methods [9] [10] [11] [14] [15]. In elbow method, the optimum number of clusters is found by minimizing the total sum of squared distances between the points and the centroids. To illustrate the method, consider the data points and the clusters shown in **Figure 4**. For each data point, the distance (*i.e.*, d_i) between the data point and the centroid of the corresponding cluster (*i.e.*, cluster that the point lies) is computed. Subsequently, the total sum of squared distance between the points and the centroids (*i.e.*, $\sum d_i$) is computed.

On the other hand, silhouette method is based on the quality of the clusters. The quality of the clusters is defined based on how well the data points fall within the clusters. To demonstrate this method, consider the data point, point-A in **Figure 5**. At first, the average distance (say d_1) between the points in the cluster (*i.e.*, Cluster-1) and point-A is computed. Subsequently, the average distance between point-A and the points that fall within Cluster-2 is determined. This same procedure is repeated for Cluster-3 as well, to determine the minimum average distance between point-A and the clusters that do not contain point-A (say d_2).

Having computed these values, the quality of point-A within Cluster-1 is determined based on the silhouette value (*i.e.*, $\frac{d_2 - d_1}{\max(d_2, d_1)}$, where $\max(d_2, d_1)$ is the maximum between d_2 and d_1). This same procedure is carried out for the remining points as well. Subsequently, the average silhouette value is computed to measure the quality of the clusters. Therefore, as per this method,

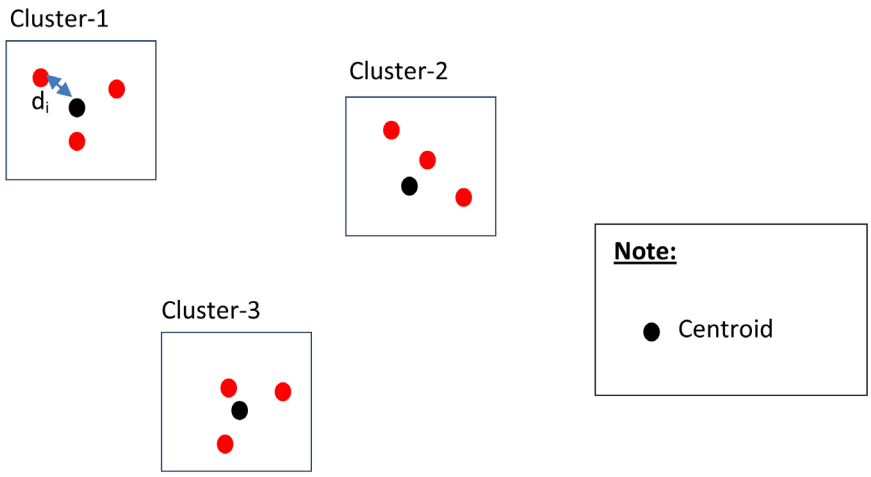


Figure 4. The elbow method to determine the optimum number of clusters.

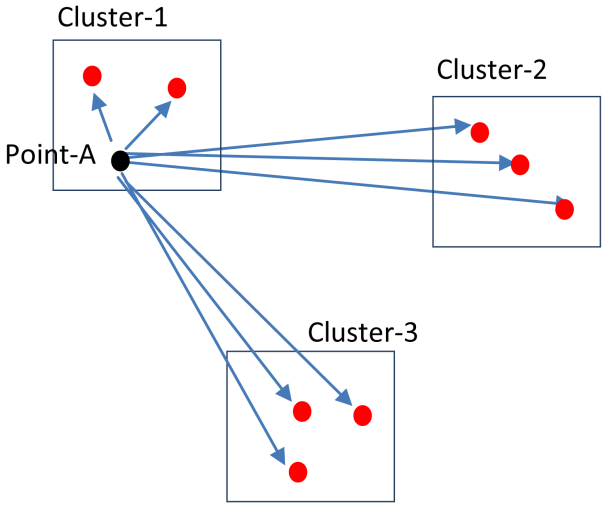


Figure 5. The silhouette method to determine the optimum number of clusters.

higher the average silhouette value is better the quality of the clusters. Moreover, the average silhouette value can vary from −1 to 1. A value of 1 indicates that a point in a cluster is far away from any of the points in any of the neighboring clusters. Therefore, the optimum number of clusters is the one having the highest average silhouette value (*i.e.*, closer to 1) [14].

The gap statistic method is based on a reference dataset generated using Monte Carlo simulation. This method compares the total intra variation for different number of clusters using the actual data against the total intra variation for different number of clusters using a reference dataset generated using Monte Carlo simulation based on the maximum and minimum values extracted from the actual data [15].

3.4. Development of Virtual Human Development Index

Having clustered the member states of the UN following the Sections 3.1-3.3, the ranks of the nations within each cluster are determined based on a composite index that is termed virtual human development index (VHDI). The composite index is based on the weighted values of the principal component scores bounded between 0 and 1. The weights are based on the proportions of total variation of the principal components. The mathematical representation of VHDI is given by Equation (13) and Equation (14).

$$PCA_i = \frac{PCA_i - PCA_{i,\min}}{PCA_{i,\max} - PCA_{i,\min}} \tag{13}$$

$$\text{VHDI} = \sum_{i=1}^{p} PCA_i * \frac{\lambda_i}{\lambda_1 + \lambda_2 + \cdots + \lambda_p} \tag{14}$$

where PCA_i is the i^{th} principal component score bounded between 0 and 1, λ_i is the variation of principal component "i", and "p" is the number of explanatory variables. $PCA_{i,\min}$ and $PCA_{i,\max}$ are the minimum and maximum values of

the i^{th} principal component score, respectively.

3.5. Implementation of VHDI Using R Statistical Software

R which is supported by the R Foundation for Statistical Computing [11] is an open source extensible programming language for statistical computing and graphics. The installation of R comes with a set of packages that are known as core packages. These core packages provide a wide variety of statistical techniques such as linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, and clustering, and graphical techniques [11]. To extend the capabilities of R, there are also many community developed packages that are available at various repositories such as the Comprehensive R Archive Network [11]. Moreover, to make the coding easier using R programming language, an integrated development environment (IDE) that is known as RStudio is also available.

In the implementation of VHDI, initially, as shown in the code snippet (see snippet-1 in Appendix-A), the UNDP dataset stored in a Microsoft Excel spreadsheet is read using read_excel () and stored in a variable named "pDataAll". This variable contains six columns of data. The first three columns contain the descriptive information (*i.e.*, latitudes, longitudes, and the names of the member states of the UN), and the last three columns contain the indices (*i.e.*, LEI, EI, and II) for the member states of the UN. Since the principal component analysis is based on the indices, a new variable named "pData" is created to store the last three columns (*i.e.*, LEI, EI, and II) extracted from "pDataAll".

To perform the principal component analysis (see snippet-2 in Appendix-A), prcomp () is called. The resultant object returned from the principal component analysis is used to extract the variance associated with each principal component. Subsequently, the proportion of total variance and the cumulative proportion of total variation explained by each principal component are computed using sum () and the cumsum (), respectively. Moreover, the eigen vectors which define the direction of the principal components and the mean adjusted principal component scores are obtained by calling the rotation and x components of the resultant object returned from the principal component analysis. To visualize (see snippet-3 in Appendix-A) the proportion of total variance and the cumulative proportion of total variation explained by each principal component, few plots are developed using plot ().

Since there are three explanatory variables (*i.e.*, LEI, EI, and II) involved in the principal component analysis, a variable named "NPCA" is used to determine the number of principal components used in the subsequent analysis using clustering algorithms (see snippet-4 in Appendix-A). Based on this variable, a variable named "pDataCluster" is formed. This variable is used in the clustering algorithm. Since the optimum number of clusters is unknown, as discussed in section 3.3.1, three methods namely elbow, silhouette, and gap statistic methods are used to determine the optimum number of clusters. The implementation of these methods using factoextra package in R is shown in snippet-4 in Appen-

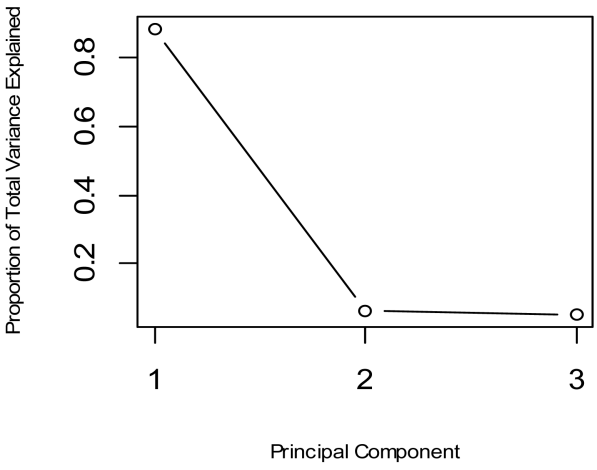
dix-A. Based on the said three methods, the optimum number of clusters is decided and set to a variable named "numberCluster". Subsequently, the k-means algorithm is called with the value of "numberCluster" (see snippet-5 in Appendix-A). The details about the clusters such as the variation within each cluster, the number of member states of the UN within each cluster, the centroids of the clusters, and the cluster number of each member states of the UN are extracted as shown in snippet-5 in Appendix-A.

To be able to visualize the clusters and to rank the member states of the UN within each cluster, a new dataset named "combinedData" is formed as shown in snippet-6 in Appendix-A. This dataset is formed based on the variable named "NPCA". For example, if the number of principal components used in the clustering algorithm is 3, the first three columns of "combinedData" contain the PCA scores bounded between 0 and 1. The fourth, fifth, and sixth columns of "combinedData" is used to store the cluster numbers, VHDIs, and the ranks of the member states of the UN within the cluster, respectively. As shown in the code snippet, initially, the sixth column is set to zero. However, this column is populated with the ranks of the member states of the UN based on the values of VHDIs within each cluster. The last few columns of "combinedData" are used to store the descriptive information about the member states of the UN (*i.e.*, latitudes, longitudes, and the names of the member states of the UN) and the indices (*i.e.*, LEI, EI, and II). These columns are added to ensure that the outcome of the clustering is visualized using Google Maps JavaScript API.

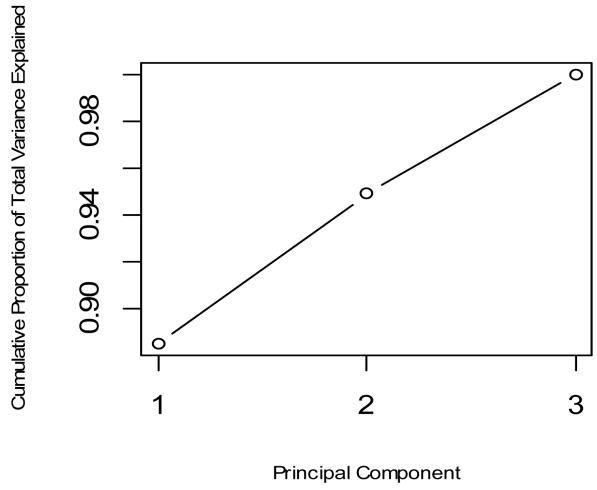
"combinedData" contains all the nations and the associated the cluster numbers. However, to rank the nations within each cluster, as shown in snippet-6 in Appendix-A, a new variable is developed to extract the nations that belong to a particular cluster number. Within each cluster, as outlined in section 3.4, the values of VHDIs are computed and stored in the variable named "combinedDataSub". As discussed previously, the fifth column which was set to zero is used to store the VHDI values. Then using the order function, the data points within a cluster are ordered and assigned the rankings for the member states of the UN within each cluster. This is accomplished using the snippet-7 in Appendix-A. Finally, as shown in snippet-8 in Appendix-A, the content of "combinedDataSub" which contains the rankings of the nations within the clusters and the associated PCA scores is exported to a Microsoft Excel spreadsheet. This Microsoft Excel spreadsheet is used to visualize the clustered member states of the UN using Google Maps JavaScript API.

4. Discussion of Results

In 2014, the proportions of total variation explained by the principal components and the cumulative proportions of total variation explained by the principal components are shown in **Figure 6(a)** and **Figure 6(b)**, respectively. The proportion of total variation explained by PCA-1 amounts to 0.885. This accounts for around 89% of the total variation. Therefore, the proportions of total



(a)



(b)

Figure 6. (a) The proportion and (b) cumulative proportion of total variance explained by the principal components in 2014.

variation explained by the remaining two principal components are very much negligible compared to the proportion of total variation explained by PCA-1. This gives an indication that among all the principal components, PCA-1 scores more towards explaining the variation. In other words, this problem of multidimensional data could be reduced to PCA-1 with little loss of information in the analysis.

4.1. Trends of Proportions of Total Eigen Value Explained by the Principal Components with Time

The proportions of total variation explained by the principal components (*i.e.*, proportions of total eigen value explained by the principal components, see Equation (10)) with time are shown in **Figure 7**. The proportion of total eigen value explained by PCA-1 increases with time though the amount of increase with time is not significant. However, the proportions of total eigen value explained

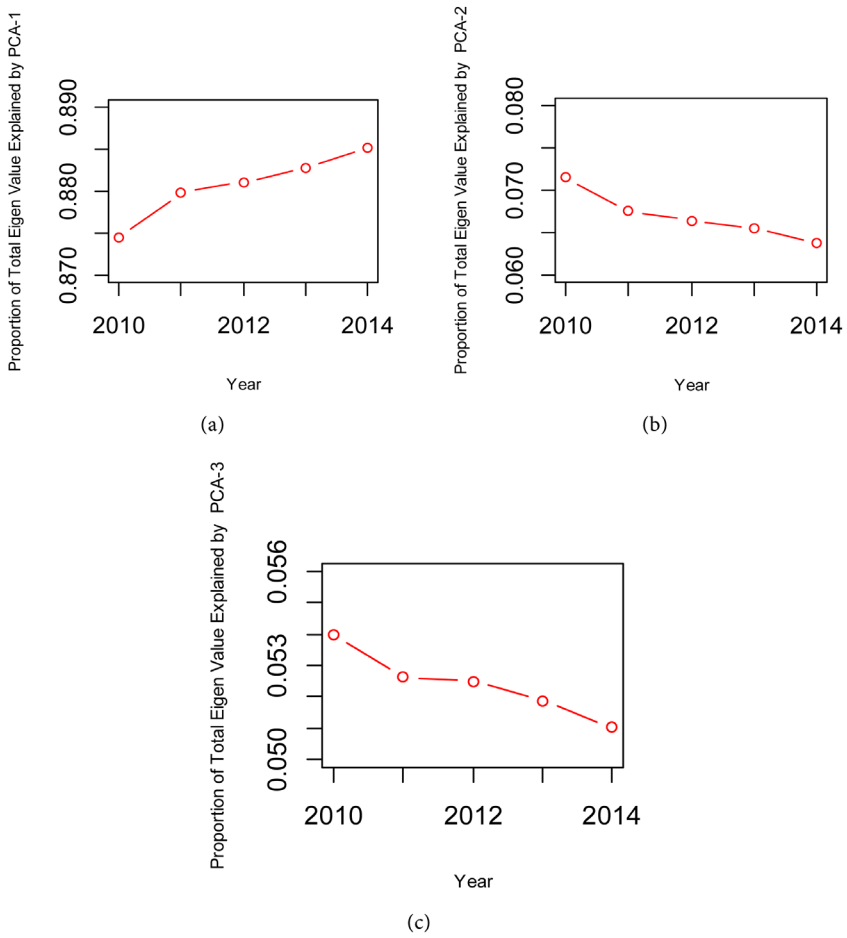


Figure 7. The proportions of total eigen value explained by PCA-1, PCA-2, and PCA-3 with Time.

by PCA-2 and PCA-3 decrease with time. Therefore, these findings indicate that, with time, the considered explanatory variables (*i.e.*, LEI, EI, and II) are well defined by choosing PCA-1 alone. In other words, the loss of information in choosing PCA-1 to represent the chosen explanatory variables may diminish with time if the trend of increasing pattern of proportion of total eigen value explained by PCA-1 with time continues in the future as well.

4.2. Correlation between the Principal Component Scores and the Explanatory Variables

To understand and interpret the principal component scores, as placed in **Table 5**, the correlation matrix between the principal component scores and the explanatory variables (*i.e.*, LEI, EI, and II) that are used in the analysis are computed using the R statistical software package. As per the computed correlation matrix, in 2014, PCA-1 is positively and very strongly correlated with all the considered explanatory variables, as depicted by the signs and the magnitudes of the correlation values. The positive correlation indicates that an increase in the value of one of the explanatory variables increases the value of PCA-1.

Table 5. The correlation matrix between the principal component scores and the explanatory variables in 2014.

	LEI	EI	II
PCA-1	0.8951269	0.9517458	0.95340788
PCA-2	−0.1012986	−0.2607833	0.29603426
PCA-3	0.4341503	−0.1617774	−0.05811306

Moreover, among all the considered explanatory variables, II which measures the standard of living has the highest correlation (0.953) followed by EI which measures the level of education and the access to knowledge (0.952), and LEI which measures the population health and the longevity (0.895) with PCA-1. On the other hand, except the correlation (0.434) between PCA-3 and LEI, the correlations between the other two principal components (*i.e.*, PCA-2 and PCA-3) and the explanatory variables are very much negligible, as depicted by the magnitudes of the correlations. Moreover, there are also negative correlations. This indicates that some of the explanatory variables have a negative trend with PCA-2 and PCA-3. Therefore, in essence, as stated previously, this problem of interest could be reduced to one principal component (*i.e.*, PCA-1) with a little loss of information in the analysis.

4.2.1. Eigen Vectors of the Principal Components

Since PCA-1 is positively and very strongly correlated with all the considered explanatory variables, as depicted by the signs and the magnitudes of the correlation values, the computed eigen vectors are evaluated. The eigen vectors that are used to weigh the explanatory variables to compute the principal scores are placed in Table 6. For PCA-1, in 2014, the coefficients that are used to weigh the contributions of EI and II are high compared to the coefficient that is used to determine the contribution of LEI. This agrees with the correlation values placed in Table 5.

4.3. Trends of Correlation between PCA-1 and the Explanatory Variables with Time

Since the proportion of total variation explained by PCA-1 in 2014 was around 89%, to understand the trends of correlation between PCA-1 and the explanatory variables (*i.e.*, LEI, EI, and II) with time, few graphs were produced as shown in Figure 8. The results show that there exists a very strong relationship between the response (*i.e.*, correlation between PCA-1 and the explanatory variable) and the explanatory variable (*i.e.*, year). The correlation between PCA-1 and EI increases with time (*i.e.*, yearly) although the magnitude of increase is not that significant. This same trend is observed in II as well. However, in contrast to these observations, LEI shows a negative trend. In other words, the correlation between PCA-1 and LEI decreases with time (*i.e.*, yearly), in addition to the fact that LEI has the lowest correlation among the considered explanatory variables

Table 6. The eigen vectors of the principal component scores in 2014.

	LEI	EI	II
PCA-1	0.4361123	0.6271851	0.6453254
PCA-2	−0.1837779	−0.6399271	0.7461360
PCA-3	0.8809267	−0.4439956	−0.1638173

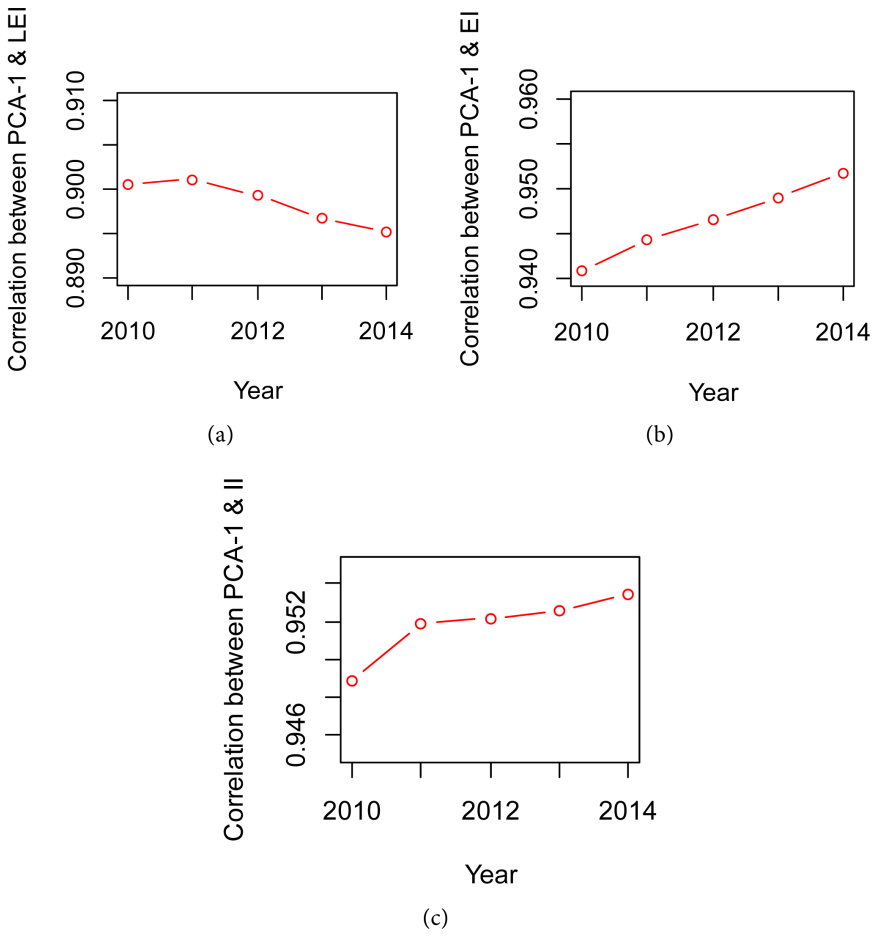


Figure 8. The Correlation between PCA-1 and the explanatory variables (*i.e.*, LEI, EI, and II).

with PCA-1. These findings imply that the contributions of EI and II to PCA-1 increase with time, but the contribution of LEI to PCA-1 decreases with time.

4.4. Determination of Clusterability and Optimum Number of Clusters

The assessment of clustering tendency which measures the clusterability was evaluated using the Hopkins statistic, as shown in **Figure 9**. The **Figure 9(a)-(c)** show the Hopkins statistic using the explanatory variables (*i.e.*, LEI, EI, and II) from 2010 to 2014; Hopkins statistic using the PCA scores (*i.e.*, PCA-1, PCA-2, and PCA-3) from 2010 to 2014; and Hopkins statistic using PCA-1 from 2010 to 2014, respectively. In all the cases, the computed values of Hopkins statistic are

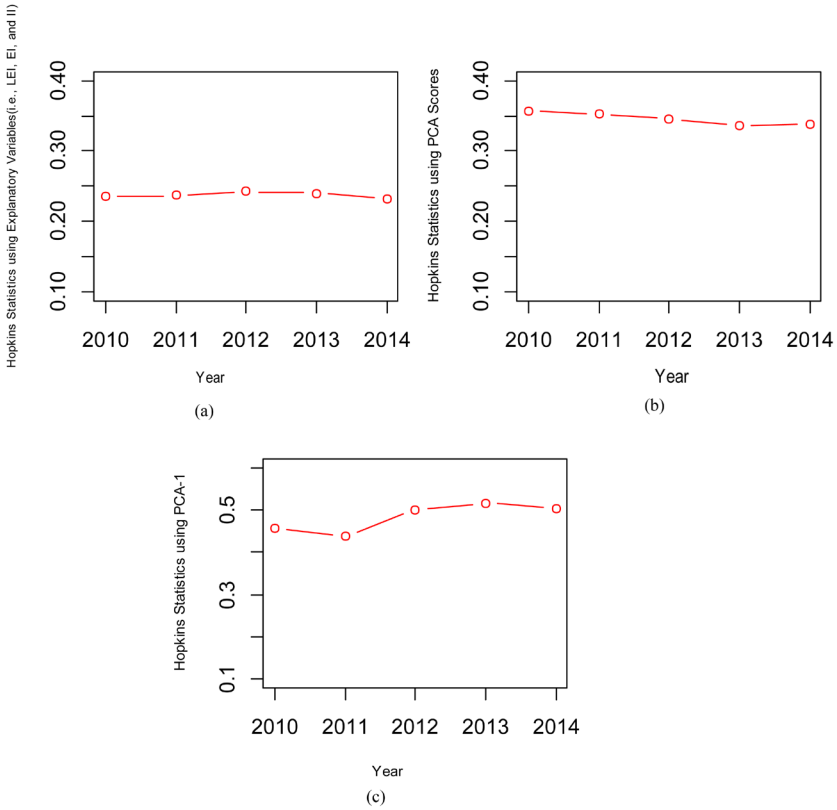


Figure 9. The trends of clusterability of PCA Scores and the explanatory variables with time.

greater than zero. Therefore, the datasets that are used to cluster the member states of the UN into tiers are not uniformly distributed and in fact clusterable. However, as portrayed by **Figure 9(a)** and **Figure 9(b)**, for the considered years, the Hopkins statistic using the PCA scores are higher than the Hopkins statistic using the explanatory variables. This is an indication that the clusterability of the information conveyed by the PCA scores is better than clusterability of the information conveyed by the explanatory variables. Moreover, as shown in **Figure 9(c)**, the clusterability of the information conveyed by PCA-1 alone is far better than the clusterability of the information conveyed by the PCA scores (*i.e.*, PCA-1, PCA-2, and PCA-3). In fact, the clusterability of the information conveyed by PCA-1 alone based on the recent data (*i.e.*, 2013 and 2014) shows Hopkins statistic reaching 0.5 and above to indicate that the clusterability of the information conveyed by PCA-1 using the recent data is better than the data from previous years.

4.4.1. Determination of Optimum Number of Clusters

The outcome of elbow method used to determine the optimum number of clusters (k) is shown in **Figure 10**. As can be observed from the figure, the total sum of squared distances between the points and the corresponding centroids (*i.e.*, intra cluster variation) decreases with increased number of clusters. However, there is a drastic drop from k = 1 to k = 2. In other words, the total sum of

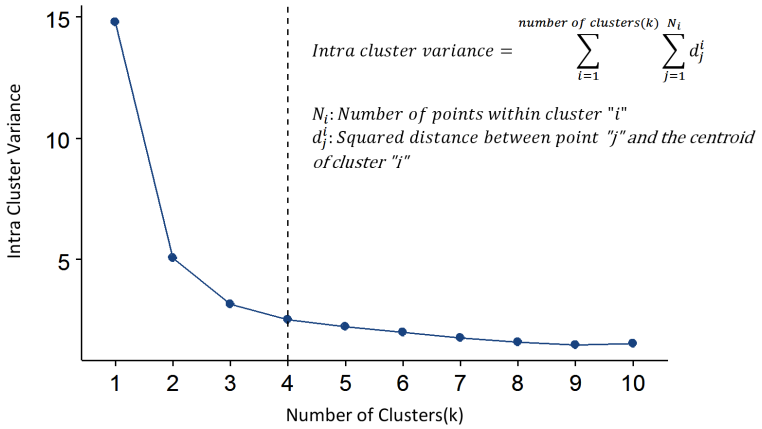


Figure 10. The optimum number of clusters using elbow method.

squared distances when k = 1 and k = 2 are 14.80743 and 5.09914, respectively. Therefore, the drop amounts to around 66%. This is an indication that there exist few tiers to pool the member states of the UN. This is also in agreement with the Hopkins statistic that measures the clusterability.

However, beyond k = 4, the total sum of squared distances between the points and the corresponding centroids does not vary much. When k = 4 and k = 5, the total sum of squared distances is 2.508759 and 2.221448, respectively. Therefore, there is not much variation between k = 4 and k = 5. This gives an indication on the number of clusters that is required to pool the member states of the UN. To investigate further on the compactness of the individual clusters when k = 4, the individual sum of squared distances between the points and the corresponding centroids are computed and placed in **Table 7**. The % total variance ((*i.e.*, intra cluster variance) within each cluster is evenly distributed among the clusters. Therefore, pooling the member states of the UN into four clusters is reasonable.

The optimum number of clusters obtained from silhouette and the gap statistic methods are placed in **Figure 11** and **Figure 12**, respectively. In silhouette method, the average silhouette value peaks when k = 2 and then decreases with k to stabilize beyond a certain value of k. Moreover, the average silhouette values computed for different ks are less than one but greater than zero. Therefore, the magnitudes of the values indicate that a point within a cluster is reasonably distanced from the points within the remaining clusters. However, since the optimum number of clusters obtained from silhouette method is two, for the same number of clusters (*i.e.*, k = 2), the compactness of individual clusters was evaluated using elbow method. The computed values are placed in **Table 8**. In addition to the fact that the intra cluster variance when k = 2 is higher compared to k = 4, the % total variance within each cluster is also not evenly distributed among the clusters. In other words, the variance within Cluster-2 is 50% more than the variance within Cluster-1. Therefore, pooling the member states of the UN into two clusters is not considered. Moreover, as per the gap statistic method which is based on a reference dataset, the optimum number of clusters is 4. This is in

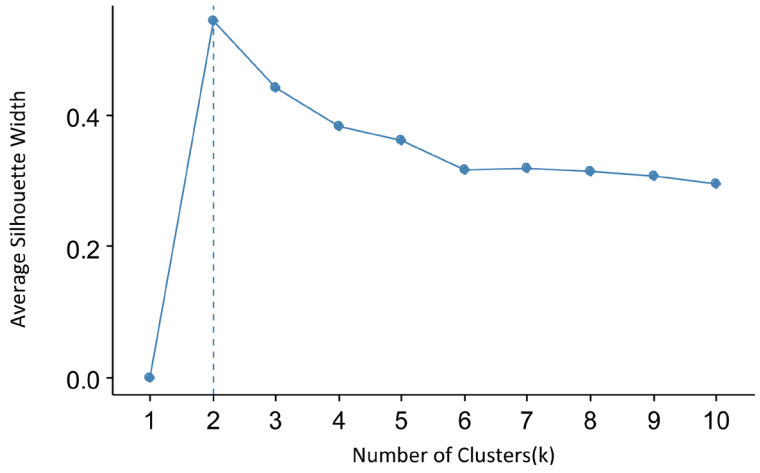


Figure 11. The optimum number of clusters using silhouette method.

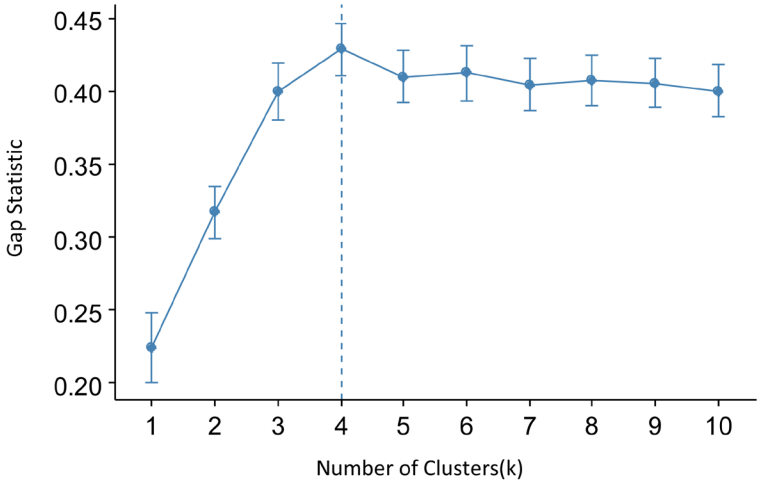


Figure 12. The optimum number of clusters using gap statistic method.

Table 7. Percentage of total variance within the clusters when k = 4.

	Cluster-1	Cluster-2	Cluster-3	Cluster-4
Total Sum of Squared Distances between the Points and the Centroid of the Cluster	0.5618820	0.8312637	0.6016126	0.5140004
% Total Variance (*i.e.*, intra cluster variance)	$=\frac{0.5618820}{2.508759}*100$ $=22.40\%$	33.13%	23.98%	20.49%

Table 8. Percentage of total variance within the clusters when k = 2.

	Cluster-1	Cluster-2
Total Sum of Squared Distances between the Points and the Centroid of the Cluster	2.02599	3.073154
% Total Variance (*i.e.*, intra cluster variance)	$=\frac{2.02599}{5.099141}*100$ $=40\%$	60%

agreement with elbow method. Therefore, based on the selected three methods, pooling the member states of the UN into four clusters is reasonable and justified.

4.5. Ranking of Member States of the UN within the Clusters

The Map of VHDI using Google Maps JavaScript API, which is accessible from http://abzwater.com/undp/hdi.php, is shown in **Figure 13**. As shown in the figure, the locations of the member states of the UN are labeled with the corresponding VHDI values. On clicking a location, as shown in the figure, a window pops up to show the rank, name of the cluster, and the values of VHDI, LEI, EI, and II. Following the human development report published in 2015 [2], the clusters are named as very high human development (VHHD), high human development (HHD), medium human development (MHD), and low human development (LHD).

In 2014, the number of member states of the UN that fall within VHHD, HHD, MHD, and LHD are 49, 71, 31, and 37, respectively. However, as per the human development report published in 2015 [2], the number of member states of the UN that fall within VHHD, HHD, MHD, and LHD are 49, 56, 39, and 44, respectively. Further investigation shows that except for Montenegro which is replaced by Russian Federation, the member states of the UN that fall under the category of VHHD are found in one of the clusters with high VHDI values. To understand the reason for the elimination of Montenegro from VHHD, the reported values of the explanatory variables for Montenegro [LEI = 0.865, EI = 0.797, II = 0.754] are compared with Russian Federation [LEI = 0.771, EI = 0.816, II = 0.828]. As can be noticed, in Russian Federation, the values of EI and II are higher than in Montenegro. This could be the reason as it was showed that the correlations of EI and II with PCA-1 are high compared to LEI, in addition to the fact that the contribution of LEI to PCA-1 decreases with time.

The member states of the UN that are pooled under LHD are found to be less compared to what is reported in the human development report published in

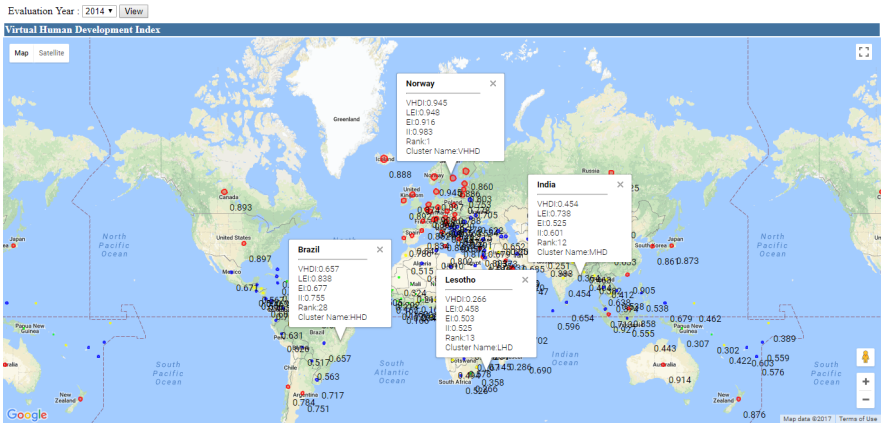


Figure 13. The map of VHDI using Google Maps JavaScript API (http://abzwater.com/undp/hdi.php).

2015 [2]. In the report, there are 44 member states of the UN in LHD. However, as per VHDI, there are 37 member states of the UN in LHD. Notably, Pakistan, Myanmar, Nepal, Swaziland, Kenya, Angola, and Nigeria are the nations that got elevated to MHD. The elevated nations are also ranked at the bottom in MHD. The likely reason for the elevation is the reason discussed in the previous paragraph.

To further understand the clustered nations, the statistical measures of VHDIs within each cluster are shown in **Figure 14**. The medians of VHDIs within VHHD, HHD, MHD, and LHD are 0.858, 0.633, 0.414, and 0.238, respectively. The magnitudes of the median values indicate that there exists a reasonable distinguishable separation between the groups or clusters. Moreover, the difference in the median values of VHDIs in between VHHD and HHD and HHD and MHD are around 0.22. However, the difference in the median values of VHDIs in between MHD and LHD is around 0.18. In other words, the possibility of elevating the member states of the UN in LHD to MHD is easier than the possibility of elevating the member states of the UN in MHD to VHD and VHD to HVHD.

Moreover, the number of member states of the UN that fall within VHHD, HHD, MHD, and LHD are around 26%, 38%, 16.5%, and 20% of the member states of the UN, respectively. Though these values may vary with time, the percentile of the member states of the UN within HHD and MHD are worth to be noted. Considering these percentile values with the medians of VHDIs discussed above, the efforts required to elevate the member states of the UN in HHD to VHHD may be sufficient to surrogate the efforts required to elevate the member states of the UN in LHD to MHD and MHD (except for few nations such as India due to its population that amounts to around 1295 million [2]) to VHD. Therefore, the possibility of having more percentile of the member states of the UN in HHD is becoming more imminent.

4.5.1. Statistical Trends of VHDIs with Time

The statistical trends of VHDIs in VHHD, HHD, MHD, and LHD are shown in

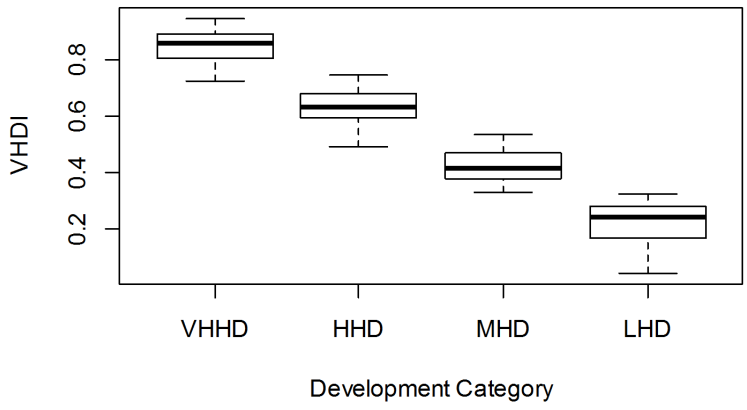


Figure 14. The statistical measures of the clusters (*i.e.*, VHHD, HHD, MHD, and LHD) in 2014.

Figures 15(a)-(d), respectively. The magnitudes of the median values of VHDIs show that the separation between the clusters or groups is distinguishable. Moreover, in VHHD, HHD, and LHD, the magnitudes of the median values have not changed significantly beyond 2012. However, in these clusters, the magnitudes of the medians are elevated compared to the previous years (*i.e.*, 2010 and 2011). In MHD, the median value of VHDI is influenced by the addition of some of the nations from LHD and the elimination of some of the top-ranking nations from MHD to VHD. Moreover, the minimum value of VHDI in HHD is decreasing with time. In 2010, the minimum value of VHDI in HHD is 0.524. However, this value has decreased to 0.492. This implies that more opportunity arises for the nations in MHD to get elevated into HHD. In other words, as mentioned previously, the possibility of having more percentile of the member states of the UN in HHD is becoming more imminent.

As per Figure 15(d), with recent data, the range between the maximum and the median values of VHDIs is narrowing. In other words, the range between the maximum and the median indicates that around 50% of the member states of the UN in LHD are close to reaching the minimum value of VHDI in MHD. Therefore, the possibility of elevating these nations that fall within this range in LHD to MHD is fast becoming feasible with proper attention. In other words, the possibility of cutting the % of the member states of the UN in LHD into 50%

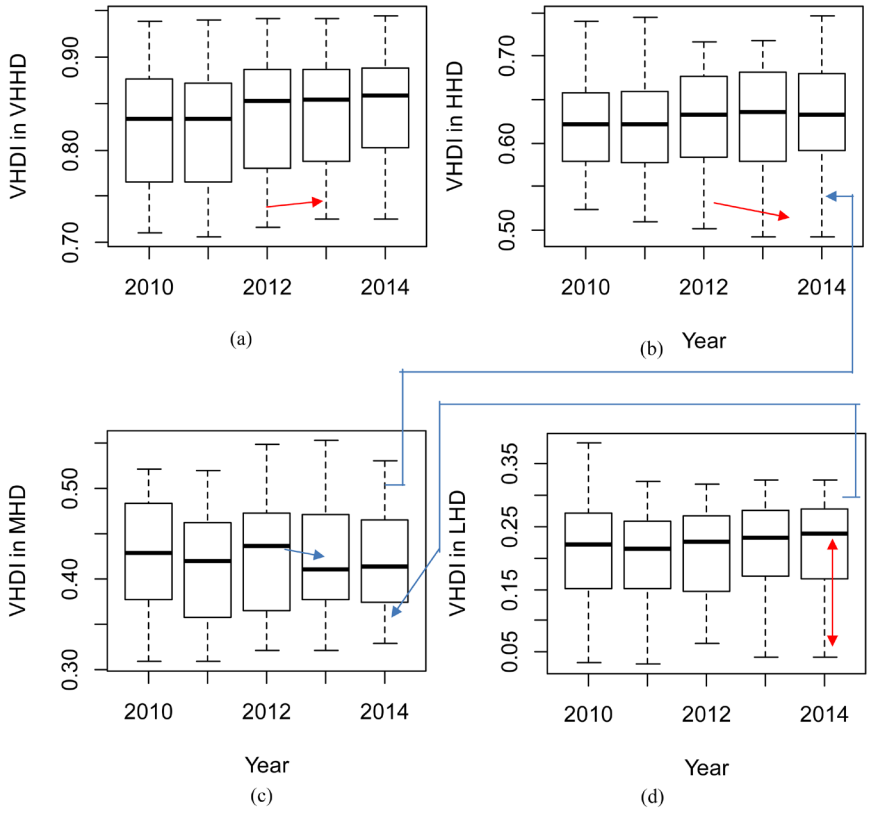


Figure 15. The statistical trends of VHDIs in (a) VHHD; (b) HHD; (c) MHD; and (d) LHD with Time.

is around the corner. However, the range between the minimum and the median value of VHDI in LHD indicates that there is a variation between the countries in this range in LHD. To worsen the matter further, as per the human development report published in 2015 [2], more than 55% of the population (*i.e.*, around 372 million out of 683 million) in LHD falls under this range. Therefore, the marginal difference among these nations in this region alarms the necessity of the attention from a global perspective.

5. Conclusions and Recommendations

In this manuscript, based on principal component analysis and K-means clustering algorithm, the data that reflect the measures of life expectancy index (LEI), education index (EI), and income index (II) are used to analyze, categorize, and rank the member states of the UN to reflect the development status. Based on this study, the following points are highlighted:

1) The proportion of total eigen value (*i.e.*, proportion of total variance) explained by PCA-1 (*i.e.*, first principal component) accounts for more than 85% of the total variation. Moreover, the proportion of total eigen value explained by PCA-1 increases with time (*i.e.*, yearly) though the amount of increase with time is not significant. However, the proportions of total eigen value explained by PCA-2 and PCA-3 decrease with time. Therefore, the loss of information in choosing PCA-1 to represent the chosen explanatory variables (*i.e.*, LEI, EI, and II) may diminish with time if the trend of increasing pattern of proportion of total eigen value explained by PCA-1 with time continues in the future as well.

2) As per the computed correlation matrix, PCA-1 is positively and very strongly correlated (correlation coefficient>0.89) with all the considered explanatory variables. Moreover, among all the considered explanatory variables, II which measures the standard of living has the highest correlation (correlation coefficient ≈ 0.95) with PCA-1 followed by EI which measures the level of education and the access to knowledge, and LEI which measures the population health and the longevity.

3) The correlation between PCA-1 and EI increases with time although the magnitude of increase is not that significant. This same trend is observed in II as well. However, in contrast to these observations, LEI shows a negative trend. In other words, the correlation between PCA-1 and LEI decreases with time. These findings imply that the contributions of EI and II to PCA-1 increase with time, but the contribution of LEI to PCA-1 decreases with time.

4) The Hopkins statistic using the PCA scores (*i.e.*, PCA-1, PCA-2, and PCA-3) is higher than the Hopkins statistic using the explanatory variables. In other words, the clusterability of the information conveyed by the PCA scores is better than clusterability of the information conveyed by the explanatory variables. However, the clusterability of the information conveyed by PCA-1 alone, specifically using the recent data (*i.e.*, 2013 and 2014), is far better than the clusterability of the information conveyed by the PCA scores. Therefore, choosing

PCA-1 to represent the chosen explanatory variables is becoming more concrete.

5) The magnitudes of the median values of Virtual Human Development Indices (VHDIs) indicate that there exists a distinguishable separation between the groups or clusters. Moreover, the possibility of elevating the member states of the UN in LHD (*i.e.*, Low Human Development) to MHD (*i.e.*, Medium Human Development) is easier than the possibility of elevating the member states of the UN in MHD to HHD (*i.e.*, High Human Development) and HHD to VHHD (*i.e.*, Very High Human Development). Moreover, the possibility of having more percentile of the member states of the UN in HHD is becoming more imminent.

6) In this manuscript, the development of VHDI is based on the reported values of LEI, EI, and II. However, it is also worth to research to fit the composite index (*i.e.*, VHDI) based on the values (*i.e.*, life expectancy at birth in years, mean years of schooling in years, expected years of schooling in years, gross domestic product per capita) that form these indices (*i.e.*, LEI, EI, and II). This will eliminate any errors which may arise in using LEI, EI, and II in the development of VHDI.

7) The VHDI is developed based on the weighted values of the principal component scores bounded between 0 and 1. The weights are based on the proportions of total variation of the principal components, considering all the member states of the UN. Therefore, it is also worth to research the impact of determining the weights, considering only the member states of the UN that fall within a cluster of interest.

8) In the definition of LEI, the maximum and minimum values (*i.e.*, goalposts) of life expectancy (LE) are set to 85 years and 20 years, respectively. The justification for setting the minimum LE at 20 years is based on historical evidence that no country in the 20th century had a life expectancy of less than 20 years [1] [2]. However, to consider the indices (*i.e.*, LEI, EI, and II) from a global perspective, these threshold values (*i.e.*, goalposts) should be based on the dataset used in the computation for the year of interest. For example, in 2015, if a maximum of 87 is observed for LE among all the member state countries, then this value should be considered as the maximum in the equation of LEI. Similarly, in 2015, if a minimum of 40 is observed for LE among all the member state countries, then this value should be considered as the minimum. This will ensure that the member state of the UN that registers the maximum value for LE is given a value of 1 for LEI. Similarly, the member state of the UN that registers the minimum value for LE is given a value of 0 for LEI. Furthermore, based on these two member states of the UN, the other member states of the UN should be evaluated on LEI. In the longer run, even if all the member states of the UN register LEs of 85 and above, the above outlined approach will sustain.

9) Though K-means clustering algorithm is the most widely used clustering algorithm, it is also worth to research on other clustering algorithms.

Acknowledgement and Disclaimer

The author is an alumnus of Texas A&M University, Texas, USA. The views ex-

pressed here are solely those of the author in his private capacity and do not in any way represent the views of Texas A&M University, Texas, USA.

References

[1] United Nations Development Programme (2016) Human Development Report 2016: Human Development for Everyone, New York, USA.

[2] United Nations Development Programme (2015) Human Development Report 2015: Work for Human Development, New York, USA.

[3] Stanton, E.A. (2007) The Human Development Index: A History. Political Economy Research Institute, Amherst (MA): University of Massachusetts, Working Paper Series no. 127.

[4] Srinivasan, T.N. (1994) Human Development: A New Paradigm or Reinvention of the Wheel? *American Economic Review*, **84**, 238-243.

[5] Ogwang, T. (1994) The Choice of Principle Variables for Computing the Human Development Index. *World Development*, **22**. https://doi.org/10.1016/0305-750X(94)90189-9

[6] Wolff, H., Chong, H. and Auffhammer, M. (2011) Classification, Detection and Consequences of Data Error: Evidence from the Human Development Index, Cornell University, School of Hospitality Administration. http://scholarship.sha.cornell.edu/articles/338

[7] Biswas, B. and Caliendo, F. (2002) A Multivariate Analysis of the Human Development Index. *Indian Economic Journal*, **49**, 96-100.

[8] Biswas, B. and Caliendo, F. (2002) A Multivariate Analysis of the Human Development Index, No 2002-11, Working Papers, Utah State University, Department of Economics.

[9] Department of Statistics Online Programs (2017) Graduate Online Courses, Pennsylvania State University, USA. https://onlinecourses.science.psu.edu/

[10] MacQueen, J. (1967) Some Methods for Classification and Analysis of Multivariate Observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, *Volume* 1: *Statistics*, University of California Press, Berkeley, Calif., 281-297. https://projecteuclid.org/euclid.bsmsp/1200512992

[11] R Core Team (2017) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/

[12] Hopkins, B. and Skellam, J.G. (1954) A New Method for Determining the Type of Distribution of Plant Individuals. *Annals of Botany*, **18**, 213-227. https://doi.org/10.1093/oxfordjournals.aob.a083391

[13] Lawson, R.G. and Jurs, P.C. (1990) New Index for Clustering Tendency and Its Application to Chemical Problems. *Journal of Chemical Information and Computer Sciences*, **30**, 36-41. https://doi.org/10.1021/ci00065a010

[14] Rousseeuw, P.J. (1987) Silhouettes: A Graphical aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics*, **20**, 53-65. https://doi.org/10.1016/0377-0427(87)90125-7

[15] Tibshirani, R., Walther, G. and Hastie, T. (2001) Estimating the Number of Data Clusters via the Gap Statistic. *Journal of the Royal Statistical Society B*, **63**, 411-423. https://doi.org/10.1111/1467-9868.00293

Appendix A

```
#Snippet-1
DataFile = "D:/Simulation_Results/HI_EI_II.xlsx"
NPCA =3
numberCluster =4
library(readxl)
pDataAll <-read_excel(DataFile)
pData <-pDataAll[, 4:ncol(pDataAll)]
pData
```

```
#Snippet-2
HDI.PCA <-prcomp(pData)
eigs <-HDI.PCA$sdev ^2
eigs /sum(eigs)

cumsum(eigs) /sum(eigs)

HDI.PCA$rotation

HDI.PCA$x
```

```
#Snippet-3
plot(eigs/sum(eigs), xlab ="Principal Component",
ylab ="Proportion of Total Variance Explained",
type ="b", xaxp    =c(1, 3, 2), cex.lab=0.7)

plot(cumsum(eigs)/sum(eigs), xlab ="Principal Component",
ylab ="Cumulative Proportion of Total Variance Explained",
type ="b", xaxp    =c(1, 3, 2), cex.lab=0.7)


#Snippet-4
pDataCluster <-cbind(HDI.PCA$x[, 1:NPCA])
pDataCluster

library(factoextra)

library(NbClust)
fviz_nbclust(pDataCluster, kmeans, method ="wss") +
geom_vline(xintercept =4, linetype =2) +
labs(subtitle ="Elbow method")

fviz_nbclust(pDataCluster, kmeans, method ="silhouette") +
labs(subtitle ="Silhouette method")

set.seed(123)
fviz_nbclust(
   pDataCluster,
   kmeans,
```

```
nstart =25,
method ="gap_stat",
nboot =50
) +
labs(subtitle ="Gap statistic method")
```

```
#Snippet-5
set.seed(7)
optimumCluster =kmeans(pDataCluster,
                                    numberCluster,
nstart =100,
algorithm ="Hartigan-Wong")
optimumCluster$size

optimumCluster$centers

optimumCluster$withinss

sum(optimumCluster$withinss)

optimumCluster$cluster

optimumCluster$totss

optimumCluster$tot.withinss

optimumCluster$betweenss
```

```
#Snippet-6
pDataClus-
ter<-cbind(PCA1=(HDI.PCA$x[,1]-min(HDI.PCA$x[,1]))/(max(HDI.PCA$x[,
1])-min(HDI.PCA$x[,1])),
PCA2=(HDI.PCA$x[,2]-min(HDI.PCA$x[,2]))/(max(HDI.PCA$x[,2])-min(H
DI.PCA$x[,2])),
PCA3=(HDI.PCA$x[,3]-min(HDI.PCA$x[,3]))/(max(HDI.PCA$x[,3])-min(H
DI.PCA$x[,3])))
pDataCluster

combinedData =cbind(pDataCluster[, 1:NPCA],
Cluster.No = optimumCluster$cluster,
VHDI =0,
VHDIR=0,
                                    pDataAll[, 1:ncol(pDataAll)])
combinedData

combinedDataSubAll=c()

for (j in 1:numberCluster) {
  combinedDataSub =subset(combinedData, combinedData[, NPCA +1] ==j)
  for (i in 1:nrow(combinedDataSub)) {
    combinedDataSub[i, NPCA +2] =sum((combinedDataSub[i, 1:NPCA])
```

```
*as.vector(eigs)[1:NPCA]/sum(eigs))

  }


#Snippet-7
combinedDataSub  =combinedDataSub[order(combinedDataSub[,  NPCA  +2],
decreasing =TRUE),]
  combinedDataSub
  for (i in 1:nrow(combinedDataSub)) {
    combinedDataSub[i, NPCA +3] =i
  }
combinedDataSub
combinedDataSubAll=rbind(combinedDataSubAll,combinedDataSub)


#Snippet-8
fileName =paste("ClusteredData", j, ".xlsx")
library(xlsx)
write.xlsx(
x =data.frame(combinedDataSub),
file = fileName,
sheetName ="Sheet1",
row.names =FALSE
  )
}
fileName =paste("ClusteredDataAll", ".xlsx")
library(xlsx)
write.xlsx(
x =data.frame(combinedDataSubAll),
file = fileName,
sheetName ="Sheet1",
row.names =FALSE
)
```