Wubetie *J Big Data* (2017) 4:47
https://doi.org/10.1186/s40537-017-0099-y

Journal of Big Data

RESEARCH Open Access

Missing data management and statistical measurement of socio-economic status: application of big data

Habtamu Tilaye Wubetie*

*Correspondence:
habtamu.tilaye@yahoo.com
Department of Statistics,
Debre Markos University,
Debre Markos, Ethiopia

Abstract

Socio-economic status measurement is an ongoing problem where different suggested measurements are given by researchers. This work investigates a socio-economic status measurement derived from natural correlations of variables which can better and meaningfully cluster African countries for the level of status. The researcher used 48 African countries socio-economic yearly time series data from 1993 to 2013 of IMF 2013 data set for data management (i.e, 2737 variables for 21 years), however, the analysis is reasonably done based on recent 14 years time series data. In data management, missing values are treated (imputed) by using regression estimates, Lagrange interpolation, linear interpolation and linear spline interpolation based on the appropriate method which best fits for the trend of data with minimum error at each time level. From principal component and factor analysis of average time series data, 7 principal factors contributed by 84 variables which explain 70% of the variation in the data set are suggested as a socio-economic status measuring components and as a result the considered clustering methods (K-mean Method, Average linkage method, Ward's method and Bootstrap Ward's method) are agreed on six clusters of countries, those are statistically significant at 95%, where as three countries each where suggested as outlier-countries made an individual cluster.

Keywords: African countries, socio-economic development, missing data management, Principal component analysis, Factor analysis and cluster analysis

Introduction

Socio-economic status measurement is an ongoing problem, where different studies had been made to measure it as a single measured variable, several single measured variables, or as a composite of several measured variables. Socio-economic status is defined as one's access to financial, social, cultural and human capital resources, and it is recommended that, family income with other indicators of home possessions and resources, parental educational attainment and parental occupational status (the "big 3") as components of a core socio-economic status measure [1]. It has been also defined as one's access to collectively desired resources , like, (1) material capital (income, wealth, trust funds, etc.), (2) human capital (skills, abilities, credentials, etc.) and (3) social capital (instrumental relationships such as being friends with lawyers and doctors) [2]. Duncan socio-economic .Index has been used in US as a measure, which is a subjective

assessment of occupational prestige based on educational attainment and income. In 1974 Peter Rossi et al. also developed a household prestige score as a measure. Currently in UK a National Statistics Socio-economic classification (NS-SEC) is used to calculate a measure for socio-economic status based on one's job and employment relations.

From sociological view status of the society can be levelled using status dimensions, those are power, wealth, prestige and information. American sociologists have used occupational prestige to level status, and they have observed for similarities or differences exist between levels at different time and place. Some sociologist have defined occupational prestige as resources availability (composed of both wealth and power) to each person where others relate it with prestige (composed of power, wealth and prestige). Goldthorpe and Hope define it as social standing which includes variables of standard of living, power and influence, level of qualification and value to society [3].

Barro [17] made a study on 100 countries from 1960 to 1990. He found that, the growth rate (real per capita GDP) is enhanced by higher initial schooling and life expectancy, lower fertility, lower government consumption, better maintenance of the rule of law, lower inflation and improvements in the terms of trade. In 2015 a study had been made on factors affecting economic growth in developing countries by using cross-country data for 76 countries from 2010, 2005, 2000 and 1995. The variables used to asses factors for GDP per capita growth are volume of export, government debt (% of GDP), natural resource yield (% of GDP), net foreign aid record (USD), life expectancy (years), Investment rate (% of GDP) and FDI inflow (% of GDP). From the result it was found that, high volume of exports, plentiful natural resources, longer life expectancy and higher investment rates have positive impacts on the growth of per capita gross domestic product in developing countries [4].

In constructing a measurement for socio-economic status, its reliability is more important. The reliability is based on socio-economic and statistical significance of the classifications made using appropriate method from a representative data. However, there is criticisms on representativeness on some of African data. The source for the problem is diversified. From history, in the 1980s and 1990s statistical offices didn't received appropriate attention as a source of data, even today the data is distorted due to the shift in data demand by donors. In addition, projects on Africa has been focusing on achieving target development rather than answering an important development questions. An other problem is lack of data, as African Development Bank survey noted nearly one-fifth of the respondent countries had not conducted an industry survey since 2000. In addition to the above problems, African data have faced sampling (inappropriate sample size and sampling technique) and non-sampling error (respondent error, non response, recording error, etc.) [5]. Missing data can be a problem when there is non-response or the data is not collected for the variable mainly not at random. Lagrange interpolation method can be used to interpolate missing values when one value is dependent on its neighbour data sets. Vaseghi [6] formulate the general form of polynomial interpolator and statistical interpolators applicable for missing data imputation purpose. He considers the special forms of Lagrange, Newton, Hermite and cubic spline interpolators for polynomial interpolators. Lokupitiya et al. [7] uses NASS data for barley crop yield in 1997 where ecological variable are spatially correlated to select a better interpolator method and find out regression and Multiple imputation as a better interpolator from

the interpolation methods considered (regression, kernel smoothing, universal kriging and multiple imputation) for Y (response) based on the target or control variable X (explanatory).

Howell [8] considered missing data problem for standard experimental studies and observational studies. In observational studies missing values can be treated by hot deck imputation, mean substitution and pairwise deletion, but those methods lead to bias in parameter estimation. However, expectation/maximization (EM) algorithm and multiple imputation (MI) are the most best techniques which are based on iterative solutions in which the parameter estimates lead to imputed values, which in turn change the parameter estimate. MI is an interesting approach because it uses randomized techniques to do its imputation an example of it is regression imputation, which regresses the response variable based on the explanatory variable.

The lack of reliable data on African countries economy limits knowledge on the economic effect of structural adjustment, as a result the economic growth time series for African economies does not appropriately capture changes in economic development [9]. Currently a better African socio-economic data is IMF [11], alternatively if we trust the AfDB, they may miss a few base year revision [10], though, AfDB is not really fully agree on IMF [11] report. Meanwhile, the AfDB, conclude that: "Overall, the situation with regard to GDP is not nearly as bad as has recently been suggested" [9]. In working over this problem, since considering the distribution of the data gives detailed and general information about the characteristics of interest than one value times series data is preferable than using single value. Consequently the risk of govern by only inappropriate data can be reduced.

Previous study on measurement of socio-economic status construct measuring components or variables based on socio-economic stand of an individual or community [1–4]. However, socio-economic status measurement is still ongoing problem. This paper desired to construct measuring components by investigating a natural correlation exist between possible suggested variables, those can able to cluster countries based on socio-economic status and level the status for components. However, a single measure for a status is not constructed, since the concern is to give specific suggestion based on the stand of cluster countries for components. Hence, time series data is used to manage the African socio-economic data problem and determine components of socio-economic status measurement which can classify African countries based on status through comparison across the region. Correspondingly, missing values were treated by a method which give minimum error from the true value at each time interval. Missing values are imputed using linear regression model, Lagrange interpolation, Linear interpolation and linear spline interpolation. Principal component analysis, factor analysis and cluster analysis are used in determining principal factor of socio-economic status measurement and clustering African countries based on those factors. The result reveals that 70% of the variation in the data set is explained by the suggested 7 components (principle factors), which are contributed by 84 variables, and using those socio-economic components 6 cluster of African countries are formed at 95% confidence level were 3 countries are consider as outlier.

Methodology

Data and variable

Data IMF [11] socio-economic yearly time series data set containing 2737 variables [File Name: 21yearData.csv] from year 1993 to 2013 for 48 African Countries were used for data management, however, the analysis is reasonably done based on the data set from 2000 to 2013. The reason of using 14 years of data instead of 21 is due to the recently growing demand for data which apparently increases outputs from statistical offices. This leads the missing value to decrease in recent years. Specifically almost all the 44 respondent countries have carried out at least one household survey of income or expenditure since 2000 [9].

The preference of data set from IMF [11] over AfDB [10] is made due to advantages listed below in (1) and (2)

1. IMF [11] a data have best coverage (48 countries) than the AfDB [10] data (44 countries).
2. Moreover, as it was indicated on the introduction section AfDB does also agree on the [11] report.

More over, Morten Jerven [9] also advises that 11. If we use AfDB they may have missed a few base year revisions.

Variables In this study the proposed components are selected based on suggested results from previous studies mainly by Cowan et al. [1], and Oakes and Rossi [2]. Cowan et al. [1] recommend that. the socio-economic status component should include family income, parental educational attainment and parental occupational status. More over, expanded measure of socio-economic status can be constructed by adding home neighbourhood and school socio-economic status. Where family income includes home possessions (internet access, clothes dryer, dishwasher, more than one bath-room, one's own bedroom), presence of household member needing healthcare assistance and household composition like size of household (total, number of adults). Correspondingly Oakes and Rossi [2] recommends material capital (income, wealth, trust funds, etc.), human capital (skills, abilities, credentials, etc.) and social capital (instrumental relationships such as being friends with lawyers and doctors). Hence, relative to the IMF data components, the proposed components are related to education, economy, health, infrastructure and population demographic data category. However, current IMF [11] African socio-economic data have three major problems:

1. Some data values are missing.
2. Different country have different base year for their GDP. In response IMF update each country's GDP based on its base year. Hence, there may be a loss in information, and comparison using single year data is inappropriate.
3. The data have some discrepancies or some davit from AfDB [10] data [as AfDB [10] conclusion: IMF GDP report is not nearly as bad as has recently been suggested]. This problem is mainly raised in data collection, processing and distribution phase. It is a duty and responsibility of statistical offices or any primary data source organizations to apply appropriate data collection techniques and standardized processing

method, and honest distribution based on the nature of the data as data is the public property. To manage this problem, distribution based analysis can reduce the risk of inferences and give relevant result than using a single value. Hence, considering time series of the data can help to do this job. For instance considering the time series of the data acquires the progress of the GDP, makes the comparison more appropriate in contrast to using 1 year GDP.

Missing data management

Missing data value is the absence of the data value completely at random (if missing values of any variable dose not depend on any value) or at random (if missing values in response variable does not depend on in its' own value but dependent on other variables) or may be not at random (if missing values follow some structure or model) [8]. The data series of African socio-economic variables on fixed time-interval have high number of missing values for some countries comparing to other's. As listed below in the sequence of missing values per country, countries such as, Somalia, South Sudan, and Sao Tome and Principe have high number of missing values compared to Tunisia, Morocco and South Africa. This suggests the probability to be missed is dependent on its' own value.

List for the number of missing values per country:

```
TUN MAR ZAF CMR MUS KEN BWA SEN GHA SDN BEN NGA
280 286 314 331 335 345 347 347 353 355 362 362

UGA BFA MOZ TGO MDG DZA MLI NER MWI NAM BDI MRT
364 367 368 368 374 381 382 383 384 394 397 400

LSO RWA ZMB GIN ETH SWZ CPV GAB SLE TCD CAF DJI
406 410 416 420 427 428 433 438 443 452 466 470

ZWE SYC AGO GNB COM ERI LBY GNQ LBR STP SOM SSD
471 476 478 482 509 526 567 569 605 629 833 931
```

In addition, since each socio-economic variable is expressed in time and have strong indirect correlation ($r = -0.8413$), the recent year has less probability of having missing value than the old one. Hence, missing values are not at random and non-ignorable.

The time series of missing values:

```
year:            1993 1994 1995  1996 1997 1998
No.of Missings:1183  1154 1130  1104 1072 1098

1999 2000  2001  2002  2003  2004  2005  2006
1054  921   1030   856  1031   893   985  968

2007 2008   2009  2010   2011    2012  2013
904  935   947   883     941    902    873
```

Interpolation

Interpolation is the estimation of unknown values using the values of known samples at the neighbourhood points [6].

Interpolation by simple linear regression method

Linear regression estimate imputation is one of the single imputation method used the surviving creature characteristicswhen the variable with missing value has correlation with explanatory variable (time) and the series of data values follow linear trend [7]. However, socio-economic data expect to have some trends but may not exactly linear in time. Hence, applying this method may enhance correlation and under estimate the standard error of the regression coefficients by under estimating the variance of the imputed variables [8]. So with this consideration simple linear regression estimate is used to impute when the missing data is at the beginning (t_1) or/and at the end (t_n). However, missing values at the internal part were treated by comparing this method for minimum error with other exact estimation methods discussed in "Linear interpolation", "Linear spline interpolation" and "Lagrange polynomial interpolation" sections.

The simple linear regression model for the given response variable Y (socio-economic variable) and the explanatory variable time (t) is given by:

$$Y_i = B_{0i} + B_{1ij}t_{ij} + \varepsilon_{ij}, \quad \text{for } i = 1, 2, \ldots, 2737, \quad \text{for } j = 1, 2, \ldots, 21. \tag{1}$$

where B_0 and B_1 are intercept and slope parameters respectively, and ε_i is error term which is normally distributed with mean μ and variance σ^2, i.e, $\varepsilon_i \sim N(\mu, \sigma^2)$.

Linear interpolation

Linear interpolation is the simplest interpolation techniques for missing data imputation using the two known neighbours. For a time series of discreet data points of socio-economic variable given by $\{(t_1, y_{t_1}), (t_2, y_{t_2}), \ldots (t_n, y_{t_n})\}$, when there is missing value at t_i, for known $y_{t_{i-1}}$ and $y_{t_{i+1}}$ linear interpolation can be used to estimate y_{t_i} at t_i by interpolating based on it's neighbours $y_{t_{i-1}}$ and $y_{t_{i+1}}$ by using the following formula.

$$Y_{t_i} = Y_{t_{i-1}} + \frac{t_i - t_{i-1}}{t_{i+1} - t_{i-1}}\left(Y_{t_{i+1}} - Y_{t_{i-1}}\right). \tag{2}$$

This method is employed to interpolate non-sequential missing values. An illustration example is presented on Fig. 3 and Table 2.

Linear spline interpolation

This method works in similar fashion as linear interpolation in a way that the missing value is interpolated by using its most two neighbours except it works for sequentially missed values. Here with some adjustment (i.e., the same upper neighbour) this method is applied to interpolate sequentially missed values. For a time series of discreet data points of socio-economic status given by $\{(t_1, y_{t_1}), (t_2, y_{t_2}), \ldots (t_n, y_{t_n})\}$, when the sequence of values $y_{t_2}, y_{t_3} \ldots y_{t_{n-1}}$ are missed.

Then for any k, $2 \leq k \leq n-1$, y_{t_k} is interpolated as:

$$Y_{t_k} = Y_{t_{k-1}} + \frac{t_k - t_{k-1}}{t_n - t_{k-1}}\left(Y_{t_n} - Y_{t_{k-1}}\right). \tag{3}$$

That is,

$$
\begin{aligned}
Y_{t_2} &= Y_{t_1} + \frac{t_2 - t_1}{t_n - t_1}\left(Y_{t_n} - Y_{t_1}\right)\\
Y_{t_3} &= Y_{t_2} + \frac{t_3 - t_2}{t_n - t_2}\left(Y_{t_n} - Y_{t_2}\right)\\
&\vdots\\
Y_{t_{n-1}} &= Y_{t_{n-2}} + \frac{t_{n-1} - t_{n-2}}{t_n - t_{n-2}}\left(Y_{t_n} - Y_{t_{n-2}}\right).
\end{aligned}
$$

On this paper linear interpolation is employed to interpolate sequentially missed values. An illustration example is presented on Fig. 4 and Table 3.

Lagrange polynomial interpolation

Lagrange polynomial interpolation is one type of exact interpolation which uses all given neighbours to estimate missing values.

For $Y_{t_i} = f(t_i)$, where, $\{t_1 < t_2 < \cdots\}$: is the function given at discreet time for socio-economic variable given by: $\left\{\left(t_1, y_{t_1}\right), \left(t_2, y_{t_2}\right), \ldots \left(t_n, y_{t_n}\right)\right\}$. The Lagrange polynomial (the nth order polynomial) for the given points is used to approximate or estimate a function $Y_{t_i} = f(t_i)$ at any time point t_i in the range, this process is called interpolation by Lagrange polynomial. For a missing value Y_{t_i} in the series of variable values the Lagrangian estimate was calculated by the following equation [6].

$$
Y_{t_i} = \sum_{k=1}^{n} \prod_{\substack{j=1 \\ j \neq k}}^{n} \left(\frac{t_i - t_j}{t_k - t_j}\right) Y_{t_k}. \tag{4}
$$

On this paper Lagrange interpolation is applied to interpolate when there is only one missing value in variable values or left with one missing value after other methods are employed. An illustration example is presented on Fig. 2 and Table 1.

Appreciatively, from theoretical advantage of Lagrange polynomial interpolation, since this method considers all known data value of a variable to estimate missing value at the point, the estimate is not only governed by its two most neighbouring data. However, due to the complication of the formula this method is employed when there is only one missing value in variable values or left with one missing value after other methods are employed.

Normality assumption It is known that the surviving creature characteristics is normally distributed. As usual, in our case this assumption is important to infer for a population because socio-economic status of African population and the variables that determine these characteristics are expected to be normally distributed. Moreover, from central limit theorem, we have the property that, the sampling distribution of the sample statistic approaches to normal distribution as sample size increases ($n > 30$) and from law of large number, we have the property that, as sample size increases the sample statistic approaches to the population parameter.

Data set of socio-economic status used for analysis is a multivariate time series data set of 2737 variables from 48 African countries for a year from 2000 to 2013.

Through the analysis of socio-economic status, it is expected that some variables have high contribution or effect on the status of socio-economic well-being comparatively. In addition, some variables may be highly correlated. Therefore, to avoid complexity due to having large number of variables, it is better to consider the possible small number of variables those can reflect the needed information. This can be done specifically:

1. When some variables are highly correlated to each other, those variables are describing the underlined characteristics which is governed by their correlation, so this characteristics will be the interest on the group. The characteristics as a new variable can be written as a linear combination of those correlated variables which can maximize the accounted variation from the total variation in the data set.
2. When some variables are correlated to the same latent or may be new variable in describing the situation of interest (socio-economic status), the latent or new variable as a linear combination of these variables is taken, in a way that the linear combination can maximize the accounted variation.
3. From the set of variables, some variables may accounted for large amount of variability in the data set. Hence, these variables can express larger amount of variation in the data set, so we can take those variables which can address the variation need to accounted.

In general the above three theories lead to principal component analysis and explanatory factor analysis.

In another word, we are assessing the variation between the random variables and variance of a variable. Normally, the variation between random variables is estimated by the distance variation of each random variable from their mean in units of standard deviation. This distance is a standardized and correlation free random variable [12]. In doing so, statistical distance plays an important role because the smaller this distance between the variables implies high correlation (it is observed on the off-diagonal of correlation or covariance matrix).

For the multivariate normally distributed random variables denoted by the random vector $\mathbf{Y}' = [Y_1, Y_2, \ldots Y_p]$, with p-dimensional normal density given by:

$$f(y) = \frac{1}{(2\pi)^{\frac{p}{2}} \mid \Sigma \mid^{\frac{1}{2}}} exp^{-\frac{(\mathbf{y}-\mu)'\Sigma^{-1}(\mathbf{y}-\mu)}{2}},$$

where p is the total number of variables, and y_i, for $i = 1, 2, \ldots, p$ (for $p = 2737$) is of an n component normal random variable with mean μ and variance σ^2.

The squared statistical distance from $\mathbf{Y}$ to population mean μ for $p \times 1$ vector $\mathbf{y}$ of observations is given by:

$$(\mathbf{y} - \mu)'\Sigma^{-1}(\mathbf{y} - \mu),$$

where the $p \times 1$ vector μ represent the expected value of the random vector $\mathbf{Y}$ and $p \times p$ matrix Σ is the variance–covariance matrix of $\mathbf{Y}$

Principal component analysis

Principal component analysis describes the correlation or variance–covariance structure between the set of variables through a few uncorrelated latent or new variables, each of which is a linear combination of the original variables which can maximize the variance accounted. Most often these new variables reveals a new interpretation that is not visible in original variables [13]. The newly created variables are called principal components.

Let the random vector $\mathbf{Y}' = [Y_1, Y_2, \ldots Y_p]$ have covariance matrix Σ with eigenvalue–eigenvector pair (λ_1, e_1), (λ_2, e_2), ..., (λ_p, e_p), where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. Then ith principal component Z_i for $i = 1, 2, \ldots, k$ where $k \leq p$ is a linear combination given by:

$$Z_i = \mathbf{e}_i'\mathbf{Y} = e_{i1}Y_1 + e_{i2}Y_2 \ldots e_{ip}Y_p$$

with $Var(Z_i) = \mathbf{e}_i'\Sigma\mathbf{e}_i = \lambda_i$ for $i = 1, 2, \ldots, k$ and $Cov(Z_i, Z_j) = \mathbf{e}_i'\Sigma\mathbf{e}_j = 0$ for $i \neq j$ which maximizes $Var(Z_i) = \mathbf{e}_i'\Sigma\mathbf{e}_i$.

Principal components are arranged in decreasing order based on the proportion of the variation they can explain, in away that the first principal component accounts for the maximum variation than any of others. Therefore, taking the most first components may can address most of the variation in the original data (like up to 80 or 90% of the population variation). However, deciding the number of principal components have no yet well stated rule, but it is also advisable to consider the size of eigenvalues and the nature of components. Most often principal components with relatively equivalent small size eigenvalues are not consider. In general, it helps to reduce the data size in variable and shows the correlation between the variables on it [12], and those new variables are used for further analysis like cluster and regression analysis.

Factor analysis

Factor analysis is used to describe the observed correlation (covariance relation) between the variables in terms of few new random variables called factors [13]. This method concerns about grouping highly correlated variables together in a way that variables in different groups are relatively slightly correlated. So in a group, those variables are addressing a characteristics which is governed by the underline correlation, called factor. Factors are sightly correlated new variables.

For the random vector $\mathbf{Y}' = [Y_1, Y_2, \ldots Y_p]$ with mean vector μ and covariance matrix Σ. The factor model postulates that $\mathbf{Y}$ is linearly dependent on a $k \times 1$ random vector $\mathbf{F}$ called common factors and a $p \times p$ diagonal matrix ε called specific factors. Then the interrelation between the elements of $\mathbf{Y}$ is given by a factor model:

$$\mathbf{Y} = \mu + \Lambda\mathbf{F} + \varepsilon,$$

where Λ is $p \times k$ matrix of unknown constants called loadings.

Assumptions of factor model on $\mathbf{F}$ and ε;

1. $\mathbf{F} \sim N(\mathbf{0}, \mathbf{I})$.
2. $\varepsilon \sim N(\mathbf{0}, \Phi)$, where $\Phi = diag(\phi_1, \phi_2, \ldots, \phi_p)$.
3. $\mathbf{F}$ and ε are independent. This assumption leads us to estimate covariance matrix, given by:

$$\Sigma = \Lambda\Lambda' + \Phi,$$

and, $Cov(\mathbf{F}, \mathbf{Y}) = \mathbf{L}$ or $Cov(Y_i, F_j) = l_{ij}$,

where $h_i^2 = \sum_{j=1}^{k} l_{ij}^2$ (Communality) and $\phi_i = Var(Y_i) - h_i^2$ (Uniqueness), for $i = 1, 2, 3, \ldots p$

The comparison of estimate of covariance to the original covariance tells us how the factor model fits the covariance matrix of original variable by the considered factors. Minimum discrepancy shows the good fit. Moreover, communality and uniqueness tell us the variance accounted by factors. Specifically the *i*th communality tells us the portion of the variance of Y_i explained by k common factors and *i*th uniqueness tells about the portion of variance of Y ($Var(Y_i)$) explained by the *i*th specific factors. Our concern is mainly looking at the factor model that explains covariance structure without much loss of information by small number of common factors.

Cluster analysis

Cluster analysis is a method of grouping of objects or variables based on similarity or distance by considering the nature of the variable or scale of measurements and the subject matter knowledge in-order to make objects in a group to be similar and objects in different groups be relatively different. Usually objects, units or cases are clustered based on sort of distance, whereas variables are clustered based on correlation coefficients with a goal to find optimal group [14].

In this paper the combined method of Hierarchical Clustering followed by Non-Hierarchical Clustering including bootstrap Ward's method were used due to the advantages of Hierarchical Clustering is better in finding the number of groups and initial cluster members where as Non-hierarchical Clustering gives more accurate members based on initial cluster members given by hierarchical method.

1. *Hierarchical clustering method:* Hierarchical clustering is an unsupervised method of grouping list of items through successive merging based on similarity or successive division based on dissimilarity. This method fall into two categories, Agglomerative hierarchical method and Divisive hierarchical method. Divisive hierarchical method start with group of items and continues by dividing the group into two subgroups by taking most similar items together in one group till each individual item make its own cluster where as Agglomerative hierarchical method start with a single item and merge most similar items together as a group, and these groups are merged successively based on similarity until the similarity is low. Then, those groups with low similarity are taken as clusters. The choice of similarity between groups or items can be measured based on average linkage or nearest neighbour linkage or the farthest neighbour linkage between the points of the groups or ward's method. However, Agglomerative hierarchical algorithm is faster due to its computational efficiency (running time complexity $O(n^3)$) than divisive clustering algorithm (running time complexity $O(2^n)$) [15, 16]. Hence, Agglomerative hierarchical method specifically average linkage (average euclidean distance) and Ward's similarity measure are used. As described on the introduction section African socio-economic data have a problem, so working based on the characteristics of the distribution can give rel-

evant information. Hence, Average linkage helps to control the impact of a single value, so the result will not be fully affected by a probably-misleading nearest (due to single linkage method) or farthest point (due to complete linkage method). e.g. End points of Chaining cluster. The result of Agglomerative hierarchical clustering can be presented by two dimensional graphs called dendrogram or by the 95% confidence bounded ellipse scatter plot of the first and the second principal factors (which shows the proportion of variance in the data set explained by the first two components in determining clusters).

2. *Ward's hierarchical clustering and it's bootstrap extension*: In this approach the focus is minimizing the information lost due to clustering. It is clear that joining dissimilar clusters results in inflated error sum of square (ESS) and leads to much information loss. Hence, a merging with smallest change in ESS results in minimum loss of information. At the beginning each item is considered as a cluster and ESS of the i cluster is zero (ESS_i, for i= 1, 2, ..., K) and ESS of the data set is $\sum_{i=1}^{K} ESS_i = 0$, in general if there are L clusters, $ESS = ESS_1 + ESS_2 + \cdots + ESS_L$, and finally if all clusters are in one group, error sum of square is given as;

$$ESS = \sum_{i=1}^{K} (y_i - \bar{y})'(y_i - \bar{y}),$$

where y_i is the multivariate measurement associated with the ith item and $\bar{y}$ is mean of all items. The result multivariate clustering is expected to be roughly elliptical [13]. Now the equation is in how much confidence a cluster can include the items assigned by ward's method or how assigned elements of a cluster are variable. This can be check by creating a dataset using re-sampling (re-sampling may be from empirical distribution of the data or by re-sampling with replacement from the data) and do clustering for each dataset, if the proportion of an item included in the same cluster is grater or equal to the desired level of confidence, then an item is assigned to the cluster in the given confidence level.

3. *Non-hierarchical clustering method:* Non-hierarchical clustering techniques are designed to group items, rather than variables, into a collection of K clusters, which is predetermined in our case by hierarchical clustering techniques [13]. Non-hierarchical clustering is started either from random partitioning of items into K initial clusters or an initial set point which will form clusters. This paper uses the popular Non-hierarchical clustering method called K-mean method, which starts by random partitioning of items into K initial clusters and goes through the list of items for assigning an item to a cluster with a closest mean to an item.

Numerical examples for missing data management

The following examples are the realization of Lagrange interpolation, linear spline interpolation, linear interpolation and linear regression estimation for artificially made missed value/s from the known data values of one of the IMF [11] data set variables, in case of Exports of goods and services (% of GDP) for Algeria. This examples are also used to illustrate and compare the error trends made by each method in interpolating or estimating artificially missing values (Fig. 1).

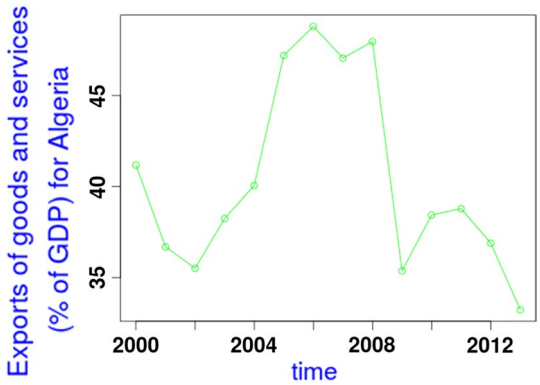

Fig. 1 Time series plot for 'Exports of goods and services (% of GDP)' for Algeria with mean 40.39 and standard deviation 5.24

Example 1 Estimation of artificially missed value for the known series of data (Y_ Actual), when a missing observation is at any point between the first and the last observation of the variable values. The result is given in Table 1 and Fig. 2.

Example 2 Estimation of two non-sequentially artificially missing values for the known series of data (Y_ Actual), when the missing observations are at any point between the first and the last observation. Table 2 and Fig. 3 shows the estimates of missing values by linear interpolation, linear spline interpolation and regression estimation methods. Here the considered cases are, when the first missing observation is at position i the second is at position $i + 4$, for $i = 2,3, ..., n - 5$.

Example 3 Estimation of four sequentially artificially missing values for the known series of data (Y_ Actual), when the missing observations are sequential at any interval

Table 1 Result for the estimates of a missing value by Lagrange interpolation, linear interpolation and linear regression estimation: in case of one missing value

Year	2000	2001	2002	2003	2004	2005	2006
Y(t)_Actual	41.18	36.69	35.50	38.25	40.05	47.21	48.81
Lagrange interpolation	41.18	85.34	30.58	39.45	42.07	46.31	46.92
Linear interpolation	41.18	38.34	37.47	37.78	42.73	44.43	47.14
Linear spline interpolation	41.18	42.75	42.42	41.51	40.99	40.12	39.84
Lagrange error	0.00	− 48.65	4.92	− 1.20	− 2.02	0.90	1.89
Linear interpolation error	0.00	− 1.65	− 1.96	0.47	− 2.67	2.77	1.67
Linear regression error	0.00	− 6.06	− 6.91	− 3.27	− 0.94	7.09	8.97
Year	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**
Y(t)_Actual	47.07	47.97	35.37	38.44	38.79	36.89	33.22
Lagrange interpolation	49.51	41.77	44.78	30.86	52.39	− 31.64	33.22
Linear interpolation	48.39	41.22	43.21	37.08	37.67	36.00	33.22
Linear spline interpolation	39.76	39.38	40.37	39.85	39.60	39.87	33.22
Lagrange error	− 2.45	6.20	− 9.40	7.59	− 13.61	5.25	0.00
Linear interpolation error	− 1.32	6.75	− 7.84	1.36	1.12	0.89	0.00
Linear regression error	7.31	8.59	− 5.00	− 1.40	− 0.81	− 2.98	0.00

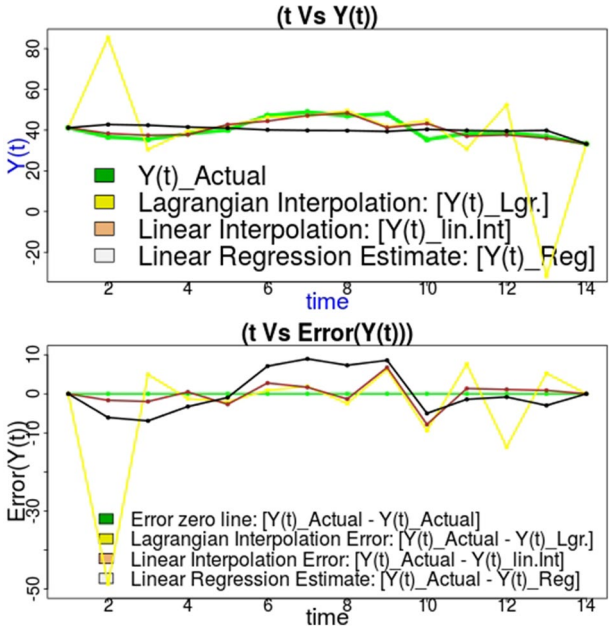

Fig. 2 The first plot is for estimates of a missing value; the second plot is for the errors in estimating a missing value

Table 2 Result for the estimates of a missing values by linear interpolation, linear spline interpolation and linear regression estimation: in case of two non-sequentially missing values

Year	2000	2001	2002	2003	2004	2005	2006
Y(t)_Actual	41.18	36.69	35.50	38.25	40.05	47.21	48.81
Linear interpolation	41.18	38.34	37.47	37.78	42.73	44.43	47.14
Linear regression estimate	41.18	41.88	41.60	41.00	40.47	40.39	39.89
Linear spline interpolation	41.18	38.34	37.47	37.78	42.73	44.43	47.14
Linear interpolation error	0.00	− 1.65	− 1.96	0.47	− 2.67	2.77	1.67
Linear Reqression error	0.00	− 5.19	− 6.10	− 2.75	− 0.42	6.81	8.92
Linear spline interpolation error	0.00	− 1.65	− 1.96	0.47	− 2.67	2.77	1.67
Year	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**
Y(t)_Actual	47.07	47.97	35.37	38.44	38.79	36.89	33.22
Linear interpolation	48.39	41.22	43.21	37.08	37.67	36.00	33.22
Linear regression estimate	39.77	39.60	39.96	39.20	38.89	38.74	33.22
Linear spline interpolation	48.39	41.22	43.21	37.08	37.67	36.00	33.22
Linear interpolation error	− 1.32	6.75	− 7.84	1.36	1.12	0.89	0.00
Linear Reqression error	7.30	8.37	− 4.58	− 0.75	− 0.10	− 1.85	0.00
Linear spline interpolation error	− 1.32	6.75	− 7.84	1.36	1.12	0.89	0.00

between the first and the last observation. Table 3 and Fig. 4 shows the estimates of missing values by linear spline interpolation and linear regression estimation methods. Here the considered cases are, when the first missing observation is at position i then the missing observation will be sequential up to $i + 3$, for i = 2,3, ..., n − 4.

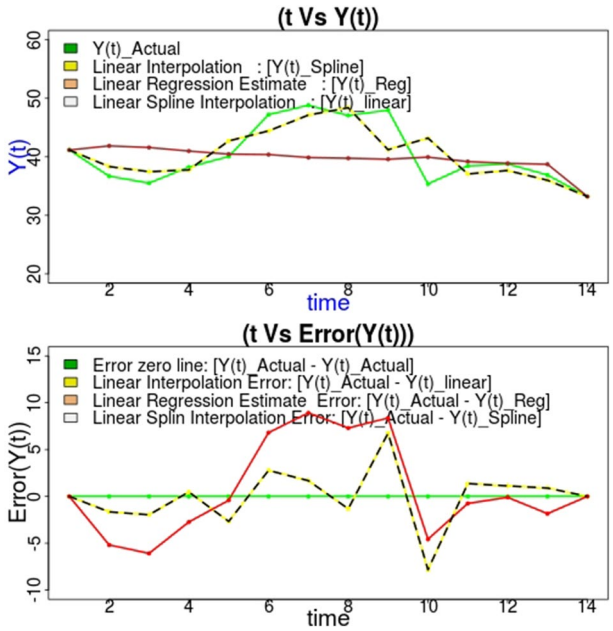

Fig. 3 The first plot for estimates of missing values; the second plot for the errors in estimating missing values

Table 3 Result for the estimates of missing values by linear spline interpolation and linear regression estimation: In case of 4 sequentially missing values

Year	2000	2001	2002	2003	2004	2005	2006
Y(t)_Actual	41.18	36.69	35.50	38.25	40.05	47.21	48.81
Linear spline interpolation	41.18	42.38	39.11	37.82	40.19	39.12	45.45
Linear regression Estimate	41.18	47.63	42.67	39.76	38.60	37.75	38.65
Linear spline interpolation error	0.00	− 5.69	− 3.61	0.43	− 0.14	8.09	3.36
Linear regression error	0.00	− 10.94	− 7.17	− 1.51	1.45	9.45	10.16
Year	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**
Y(t)_Actual	47.07	47.97	35.37	38.44	38.79	36.89	33.22
Linear spline interpolation	46.81	45.03	45.02	42.07	39.12	36.17	33.22
Linear regression Estimate	39.35	40.01	42.17	42.30	42.44	42.58	33.22
Linear spline interpolation error	0.26	2.94	− 9.65	− 3.63	− 0.33	0.72	0.00
Linear regression error	7.72	7.96	− 6.79	− 3.86	− 3.66	− 5.69	0.00

Conclusion

The above examples of interpolation methods applied for missing imputation suggests the following result.

1. Figure 3 shows that the plot of the errors due to linear interpolation and linear spline interpolation are equally closer to the horizontal error free line than the plot of the error due to linear regression line. Hence, the error due to linear interpolation and linear spline interpolation are smaller than the error due to linear regression. This reveals that, for a such time series data with non-linear trend when missing values

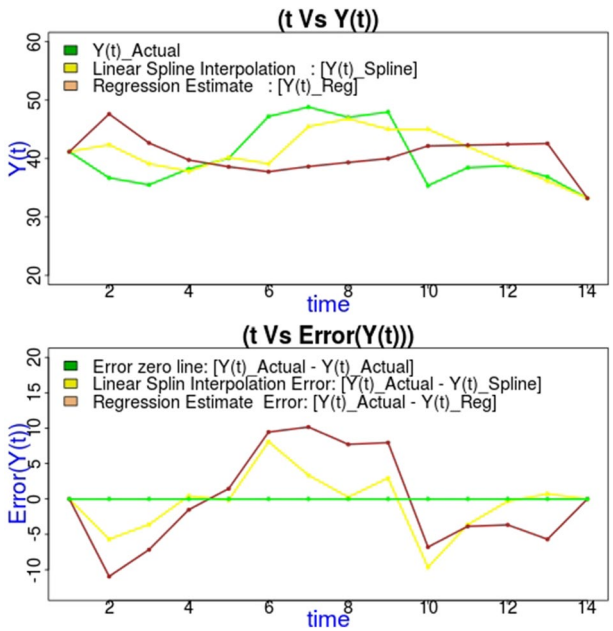

Fig. 4 The first plot is for estimates of missing values; the second plot is for the errors in estimating missing values

are not sequential, linear interpolation and linear spline interpolation brings a better estimate than linear regression.

2. Figure 4 shows that the plot of error due to linear spline interpolation is closer to the horizontal error free line than the plot of error due to linear regression. This reveals that linear spline interpolation estimator for two or more sequentially missing values have smaller error than the linear regression estimator. Therefore, for a such time series data with non-linear trend, linear spline interpolation brings a better estimate for two or more sequentially missing values than linear regression.
3. Figure 2 depicts that, since the plot of Lagrange interpolation error is closer to the horizontal zero error line than the plot of linear interpolation error and regression estimate error on the period of time interval between 5 and 9, this paper uses this method when missing values are in the middle of the observation, specifically on time interval between 5 and 9. Even-though, Lagrange polynomial interpolation gives minimum error estimator for missing value on the specified interval above, however, linear interpolation estimator have minimum error on the rest of the series comparing to Lagrange interpolation and regression estimator. Hence, linear interpolation is applied to estimate when a missing is on the other intervals (i.e, on the beginning and ending part) of the series.

Result and discussion

The concern is to formulate and apply statistical method which can grasp highly contributing variables from total variation to make major components which can significantly and meaningfully able to level the status of socio-economic development through African countries. Therefore, once the suggested socio-economic status measuring variables from literatures were considered, working on those variables by removing redundancies and

variables which have no visible role in total variance can help us to reduce the number of variables need to be considered for measurement without much loss of information. One of the techniques to do this is finding highly correlated variables (the correlation may be direct or through latent variable) and replace them by new-variable (component) which is govern by underline correlation through a linear combination of those variables, which can maximize the variance accounted by them out of the total variation. Hence, principal component and factor analysis are applied in finding key variables.

Principal component and factor analysis

This subsection considers the ways to find number of principal components needs to be considered on constructing factors and selecting highly contributor variables on total variation. The next four considerations are helping in deciding the number of principal components need to be used.

(a) *Scree plot of variance:* The scree plot in Fig. 5 shows that the bend point starts at principal factor 5. After this point the plot descends slowly but at principal components 8 there is another slight bend. Hence, two points can proposed, however, a rough view at scree plot suggests 4 principal components.

(b) *Eigenvalues:* The variance or eigenvalue of the principal components given in Table 4 reveals that, the first 14 principal components have greater than one eigenvalue, the first 8 principal components have greater than 2 eigenvalue and the first 5 components have greater than 3.76 eigenvalue. Since the eigenvalue of a principal components < 1 implies that from total variance the variance accounted by a com-

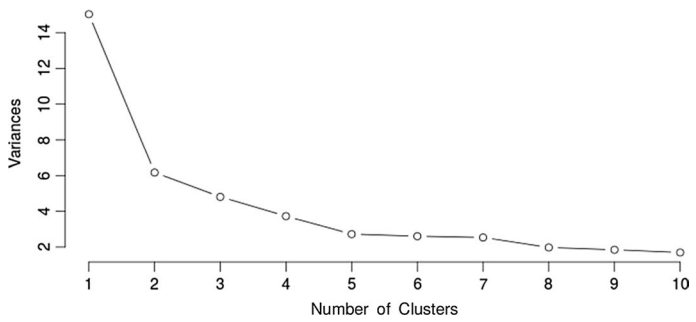

Fig. 5 A screet plot of the principal components for the average of 14 year socio-economic time series data

Table 4 Summary for variance accounted by principal components

Principal components	PCI	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Variance	15.03	6.17	4.81	3.73	2.73	2.61	2.54	1.98
Proportion of variance	0.27	0.11	0.09	0.07	0.05	0.05	0.05	0.04
Cumulative proportion	0.27	0.38	0.46	0.53	0.58	0.63	0.67	0.71
Principal components	**PC9**	**PC10**	**PC11**	**PC12**	**PC13**	**PC14**	**PC15**	**PC16**
Variance	1.85	1.70	1.35	1.13	1.12	1.10	0.93	0.77
Proportion of variance	0.03	0.03	0.02	0.02	0.02	0.02	0.02	0.01
Cumulative proportion	0.74	0.77	0.79	0.81	0.83	0.85	0.87	0.89

ponent is less than one, the principal component with large eigenvalue were chosen to explain the variation in the data set (usually with eigenvalue > 1). Based on this aspect the first four or seven factors or 13 can be taken.

(c) *Proportion of the total variation:* Here the deal is the proportion of the total variation contributed by those factors. From the result in Table 4 principal components with greater variance are selected to satisfy the proportion of total variance want to be accounted. Therefore, based on the result suggestions in (1) and (2) if the first four components are taken only 53.068% of the variation would be explained alternatively if the first seven components are taken 67.056% of the variation would be explained, where taking 13 components explain upto 83.58% of variation, but considering 13 variables are not still small and the total variation accounted by newly added 5 components is only 16.524%, in-addition the scree plot dose not support it.

(d) *Subject matter consideration:* The subject matter consideration is important to have meaningful and interpretable component for socio-economic status. From these aspect it is observed that factors resulted from analysis of 4 principal components particularly factor 3 and 4 composed of variables from different categories of data and it makes factor 3 and 4 difficult to interpret and relate with real socio-economic data categories. On the other hand factors obtained based on analysis of 7 principal components are more direct to interpret and easy to relate with categories of socio-economic data (education, economic, health, infrastructure and population demographic data). To conclude from the above four reasoning taking the first 7 principal component have relative advantage in explaining more proportion of variation (i.e. up to 67% of total variation and while in factor analysis leads the factors to explain upto 70% (Appendix 1: Table 11) of total variation in the data set), and in estimating easily and meaningfully interpretable fear number of factors.

Once the number of principal factor is determined key variables for principal factor can be selected based on loadings and their correlation with principal factor. A variable with large loadings implies that it is highly contributed by the factor and high correlation implies that the variable is highly important to determine the factor. From the result it is observed that a variable weighted with high loading by a principal factor has high correlation with it. The communality also justifies this implication.

Observing correlation between key variables of a principal factor can help to control the principal factor. So, it is important to focus on those highly correlated variables and control them firstly. The result for the correlation between key variables of each factor are given in Appendix 1: Table 12. The result of factor analysis using the first 7 principal components for correlation between principal factors and their key variables, loadings and cumulative of each key variables are given in Table 5 and reveals that:

1. The key variables in principal factor 1 are related to sustainable life measure. The result from correlation between key variables in the factor 1 suggests that:

- There is negative correlation (− 0.646) between infant mortality rate and improved sanitation facility. Hence, low sanitation can be the cause for infant mortality.
- There is good direct correlation (0.618) between life expectancy at birth and improved sanitation facility.

- There is strong direct correlation (0.812) between incidence of tuberculosis and prevalence of HIV.
- Cause of death by communicable diseases and maternal, prenatal and nutrition conditions, and cause of death, by non-communicable diseases have strong negative correlation (− 0.862), this implies that attention was given for one of them, so attention should be given for communicable diseases too.
- There is direct correlation (0.71) between infant mortality rate and communicable diseases.
- There is indirect correlation (− 0.70) between life expectancy and communicable diseases.
- Life expectancy at birth and infant mortality rate have strong negative correlation (− 0.844). This implies that most of the countries with short life expectancy should decrease infant mortality by improving sanitation problem.
- The general suggestion for the source of short life expectancy in Africa leads to low sanitation and death due to communicable diseases.

2. Principal factor 2 is related to capital. The correlation between key variables of principal factor 2 suggests the following results.

- Labour force and population is highly correlated (0.932). This is a reflection for most populated area have high labour force.
- There is strong direct correlation between transportation systems. A country with a better Air transport have a better Rail lines and Container port traffic (0.83 and 0.85 respectively), and a country with a better Rail lines have also a better Container port traffic (0.80).
- Air transport and Rail lines have strong direct correlation with GDP at market price (0.79 and 0.74 respectively). There is also strong correlation between Air transport and Gross capital formation (0.78). The correlation suggests that transportation system have strong influence on GDP at market price and Gross capital formation.
- GDP at market price have strong positive correlation with Gross capital formation and Foreign direct investment (0.925 and 0.85 respectively). Hence, GDP at market price of a country can be enhanced by calling Foreign investment and accumulating capital.
- In general Foreign investment, accumulating capital and transportation system have strong influence on GDP at market price.

3. Principal factor 3 is general income related factor. The correlation between variables of principal factor 3 suggests the following results.

- Electric power consumption have high correlation with GDP per capita, PPP (current international $) (0.753) and mobile cellular subscriptions (0.761).

- GDP per capita, PPP (current international $) and Improved sanitation facilities have strong correlation (0.744).

4. Principal factor 4 is related to life risk. The correlation between variables of principal factor 4 suggests the following results.

- Prevalence of HIV and incidence of tuberculosis have some negative correlation (− 0.415, − 0.497, respectively) with life expectancy.
- Prevalence of HIV and incidence of tuberculosis have some what visible correlation (0.442, 0.362, respectively) with manufacturing. This result is a surprising result which reflects that, manufacturing areas are suspected to be the source for medium rate of prevalence of HIV and tuberculosis. Hence, health polices should consider what have to be done in manufacturing area to reduce the prevalence of HIV and incidence of Tuberculosis.

5. Principal factor 5 is more of related to literacy. The correlation between variables of principal factor 5 suggests the following results.
- Cash surplus/deficit is strongly correlated with adult literacy rate and youth literacy rate (0.704, 0.725, respectively). Hence, illiteracy reduction plays an important role for cash surplus.
6. Principal factor 6 contrasts rate of water supply and consumption. The correlation between variables of principal factor 6 suggests the following results.

- There is an indirect Annual freshwater withdrawals in Agriculture have strong indirect correlation with Annual freshwater withdrawals in domestic (− 0.922) and Annual freshwater withdrawals in industry (− 0.794).
- There is some direct correlation (0.498) between Annual freshwater withdrawals in Domestic and Annual freshwater withdrawals in industry.

7. Principal factor 7 reflects the contrast between GDP growth rate and inflation. The correlation between variables of principal factor 7 suggests the following results.

- There is high correlation between GDP per-capita growth and inflation rate (0.954). This suggests that countries with high GDP per-capita growth should control inflation. This result agree with Barro [17] suggestion.
- There is also some direct correlation (0.40) between GDP per-capita growth and export of good and services. This result agree with Upreti [4].

Table 5 Summary table for principal factors

Variable code	Loadings	Corr	Com
Principal component I (sustainable life)			
SP.DYN.LEOO.IN	1.06	0.8079716	0.9380648
SH.DTH.NCOM.ZS	1.01	0.823052	0.7976794
SH.H2O.SAFE.RU.Z	0.78	0.7575063	0.6749383
SP.POP.1564.TO.ZS	0.73	0.8764492	0.8405396
IT.NET.USER.P2	0.67	0.7321589	0.6816921
NV.SRV.TETC.ZS	0.63	0.5728387	0.7281336
SH.STA. ACSN	0.55	0.77288	0.7755206
SH.H2O.SAFE.UR.Z	0.54	0.5288412	0.6275978
SE.PRE.ENRR	0.48	0.5518474	0.394378
IT.CEL.SETS.P2	0.47	0.7351707	0.8132745
SE.TER.ENRR	0.42	0.5128509	0.4215655
SE.SEC.ENRR	0.39	0.552263	0.4368579
SH.MED.PHYS.ZS	0.38	0.351458	0.4507551
IT.NET.BBND.P2	0.37	0.51236	0.345433
SH.DYN.AIDS.ZS	− 0.41	0.0777566	0.8047301
NV.AGR.TOTL.ZS	− 0.41	− 0.7452145	0.8178983
SH.TBS.INCD	− 0.53	− 0.0656801	0.7888127
SP.DYN.CBRT.IN	− 0.74	− 0.8915574	0.8582985
SH.STA..MMRT	− 0.82	− 0.8261328	0.725329
SH.DTH. COMMA.ZS	− 0.93	− 0.7800013	0.7496144
SP.DYN.IMRT.IN	− 0.98	− 0.8936826	0.8558577
SP.DYN.CDRT.IN	− 1.04	− 0.7554875	0.8471443
Principal component 2 (capital)			
NY.GOP.MKTP.CD	0.97	0.9517201	0.9170857
IS.RRS.TOTL.KM	0.89	0.8176941	0.7891843
BX.KLT.DINV.CD.WD	0.87	0.8659704	0.8049668
NE.GDI.TOTL.CD	0.85	0.8781617	0.8110496
IS.AIR.DPRT	0.84	0.8702821	0.8434615
IS.SHP.GOOD.TU	0.74	0.7496073	0.7518197
SP.POP.TOTL	0. 7	0.6570489	0.6306013
SL.TLF.TOTL.IN	0.66	0.6468067	0.7005713
EG.USE.ELEC.KH.P	0.45	0.5377256	0.8262583
SE.TER.ENRR	0.41	0.4621814	0.4215655
ER.H2O.FWTL.K3	0.34	0.457826	0.40111402
NE.IMP.GNFS.ZS	− 0.31	− 0.3255368	0.5675932
Principal component 3 (income related factor)			
NY.GDP.PCAP.PP.C	0.89	0.9446474	0.9213499
NY.GDP.PCAP.CD	0.85	0.9250355	0.9110524
NV.IND.TOTL.ZS	0.83	0.7730768	0.6893409
GC.DOD.TOTL.GD.Z	0.7	0.8077165	0.741978
NE.EXP.GNFS.ZS	0.59	0.7130032	0.853719
EG.USE.ELEC.KH.	0.58	0.726317	0.8262583
IT.CEL.SETS.P2	0.55	0.7422919	0.8132745
SH.STA ACSN	0.46	0.6969131	0.7755206
NE.TRD.GNFS.ZS	0.45	0.5497256	0.7607122
SH.MED.BEDS.ZS	0.45	0.4365858	0.4086245
SE.SEC.ENRR	0.35	0.5159998	0.4368579
SH.H2O.SAFE.UR.Z	− 0.31	− 0.0620451	0.6275978

Table 5 continued

Variable code	Loadings	Corr	Com
NV.AGR.TOTL.ZS	− 0.38	− 0.6669364	0.8178983
SH.XPD.TOTL.ZS	− 0.46	− 0.4721462	0.5657422
NV.SRV.TETC.ZS	− 0.5	− 0.0756577	0.7281336
Principal component 4 (life risk)			
SH.DYN.AIDS.ZS	1.05	0.7965701	0.8047301
SH.TBS.INCD	1	0.7393787	0.7888127
NV.IND.MANF.ZS	0.77	0.6267752	0.5451881
SP.DYN.CDRT.IN	0.52	0.1150051	0.8471443
IS.SHP.GOOD.TU	0.39	0.4575752	0.7518197
NE.IMP.GNFS.ZS	0.39	0.5374423	0.5675932
SE.XPD.TOTL.GD.Z	0.38	0.3719803	0.3151677
NE.TRD.GNFS.ZS	0.34	0.5519894	0.7607122
SH.XPD.TOTL.ZS	0.33	0.2408635	0.5657422
SH.H2O.SAFE.UR.Z	0.31	0.4626387	0.6275978
NV.AGR.TOTL.ZS	− 0.32	− 0.6058561	0.8178983
DT.TDS.DECT.EX.Z	− 0.37	− 0.1428698	0.2930229
SP.DYN.LEOO.IN	− 0.57	− 0.1076823	0.9380648
Principal component 5 (literacy)			
SE.ADT.1524.LT.ZS	0.87	0.8778255	0.8556102
SE.ADT.LITR.ZS	0.86	0.8747162	0.8454292
GC.BAL.CASH.GD.Z	0.77	0.7345393	0.61338
SH.MED.BEDS.ZS	0.37	0.4346196	0.4086245
SP.MTR.1519.ZS	− 0.38	− 0.3940773	0.263375
Principal component 6 (Rate of Water supply and consumption contrast)			
ER.H2O.FWDM.ZS	1	0.8787253	0.8500172
ER.H2O.FWIN.ZS	0.66	0.5969486	0.478074
NY.GNS.ICTR.ZS	0.39	0.3865365	0.1993376
DT.TDS.DECT.EX.Z	0.36	0.2825562	0.2930229
NE.IMP.GNFS.ZS	0.32	0.5611142	0.5675932
NE.TRD.GNFS.ZS	0.31	0.5644163	0.7607122
ER.H2O.FWTL.K3	− 041	− 0.4826178	0.4011402
NV.IND.MANF.ZS	− 0.43	− 0.1210153	0.5451881
ER.H2O.FWAG.ZS	− 0.99	− 0.8804037	0.8455708
Principal component 7 (GDP growth rate)			
NY.GDP.PCAP.KD.Z	0.88	0.805833	0.7268232
FP.CPI.TOTL.ZG	0.86	0.7815264	0.7317425
SH.MED.CMHW.P3	0.46	0.4290214	0.2630565
NE.EXP.GNFS.ZS	0.4	0.6200583	0.853719
SH.MED.NUMW.P3	0.4	0.4986513	0.4551595
SH.MED.PHYS.ZS	0.39	0.4168554	0.4507551
NV.SRV.TETC.ZS	0.32	0.3575785	0.7281336
SH.H2O.SAFE.UR.Z	− 0.41	− 0.3178607	0.6275978
SH.XPD.TOTL.ZS	− 0.46	− 0.4648261	0.5657422

Data quality

Before doing further analysis, it is important to know the quality and nature of the data in order to find the appropriate method and make inference. We can study the quality and nature of the data by checking for outliers and distribution type (usually normality)

respectively. Since principal factor is a linear combination of all variables with some loadings, assessing for principal factor is the reflection of assessing variables. Hence, our focus is to know what nature does principal factors have.

Q–Q is used to plot to check the normality of principal factors and T-chart to assess outliers in the data set.

From Fig. 6 Q–Q plots suggest that some of the principal factors are approximately normally distributed (i.e., principal factors in plot 2, 3, 5, 6), whereas some of them show some divergences (those are principal factors in plot 1, 4 and 7), this implies that they are not far from normal distribution. So, working with them can bring relevant inference for the population parameters.

From the result of T-chart Table 6 it is observed that countries, like, Niger (NER), South Africa (ZAF), South Sudan (SSD) have extraordinary values and Equatorial Guinea (GNQ), Libya (LBY), Swaziland (SWZ) have suspected values which have

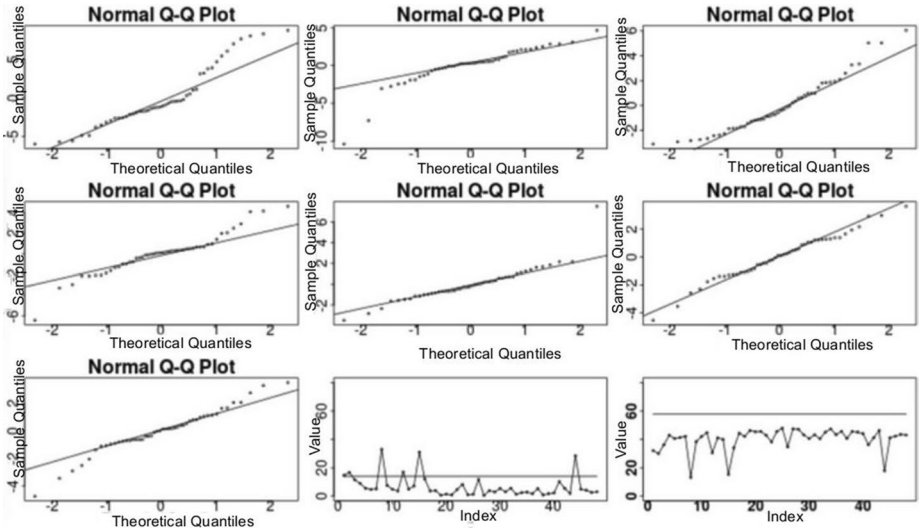

Fig. 6 A matrix plot represents: the first 7 plots for Q–Q plot of the first 7 principal components, orderly, and the last two plots (plot 8 and 9) are T-Chart plot for countries based on principal factors

Table 6 T-Chart for countries based on principal components

Country code	BENß	GHA	CMR	TGO	KEN	UGA	SEN
T-Chart	0.54	0.78	0.90	0.94	1.17	1.35	1.36
Country code	ZMB	GNB	GIN	MOZ	RWA	COM	MWI
T-Chart	1.62	1.78	1.79	1.86	2.33	2.51	2.63
Country code	BFA	TCD	SOM	MLI	MRT	ZWE	SDN
T-Chart	2.91	2.93	3.13	3.20	3.74	3.93	3.99
Country code	AGO	BDI	STP	BWA	LBR	NAM	MAR
T-Chart	3.99	4.39	4.56	4.84	4.85	4.93	5.07
Country code	ERI	SLE	DZA	MUS	MDG	CAF	CPV
T-Chart	5.22	5.36	5.46	5.63	5.65	5.71	7.39
Country code	TUN	DJI	GAB	ETH	SYC	LSO	NGA
T-Chart	7.68	8.23	8.79	10.29	11.48	11.84	12.21
Country code	GNQ	LBY	SWZ	NER	SSD	ZAF	
T-Chart	14.62	16.61	16.83	28.55	30.56	32.87	

T-chart greater than $\chi^2_{0.05,7} = 14.067$ with 95% confidence. Figure 6 plot 8 graphically shows this result. Therefore, the data from these countries need to be checked, because outliers can occurred due to coding errors, respondent errors or large true values.

Cluster analysis

A cluster analysis were used for grouping objects or variables without having any prior information or hypothesis on the number, elements and structure of the groups.

Cluster analysis is classified into two types, Hierarchical cluster analysis and Non-hierarchical cluster analysis.

Hierarchical cluster analysis is one of the preferable methods in determining the number of clusters and suggesting an initial elements of the cluster. In this analysis, average

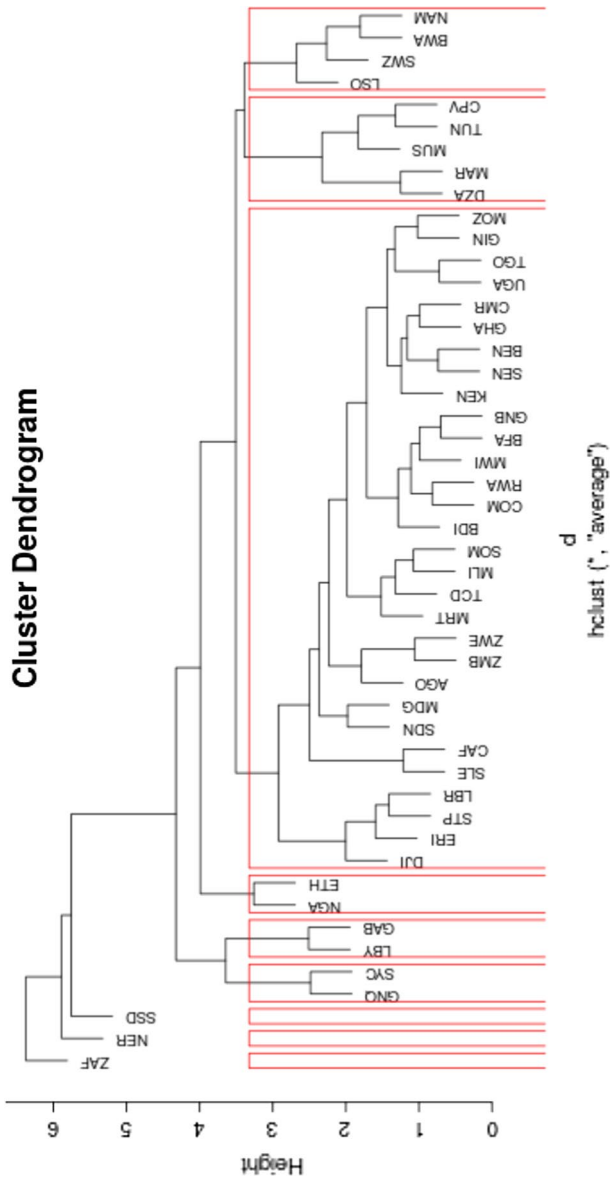

Fig. 7 Average liknage clustering of African countries based on socio-economis status

linkage, Ward's method and its bootstrap extension from the types of Agglomerative Hierarchical Cluster are used.

On the other hand Non-hierarchical clustering techniques are used to cluster countries or identify elements of the cluster based on the the number of clusters obtained from hierarchical cluster analysis. Here one of the Non-hierarchical cluster analysis called K-mean clustering is used to determine elements of the clusters in addition to the considered Agglomerative Hierarchical methods.

In this section the analysis is targeting on solving two main problems:

1. Determining appropriate number of clusters.
2. Determining elements of the cluster.

Number of clusters

The concern here is comparing the result obtained from the considered methods to estimate the number of clusters. The results from three different approaches: the Average linkage method, the scree plot of within groups sum of squares, ratio of between-cluster variability and within-cluster variability, and the Multi-scale bootstrap of Ward's method are described and discussed below in (a), (b) and (c).

(a) *Average linkage method*: From the result in dendrogram Fig. 7 based on the distance of clusters are joining, the suggestion would be 9 clusters, where three countries (South Africa, South Sudan and Niger) are each of them forming an individual cluster. Where the more the shorter distance of joining implies the more clusters are similar.

(b) *Clustering based on screen plot of within-cluster sum of square and* F_{ratio}: One of the assumption in clustering is that between-cluster variability should be relatively larger than within-cluster variability. Hence, for the given degree of freedom comparing the empirical F_{ratio} with the theoretical $F_{statistics}$ helps in decision making process of statistically significance minimum number of clusters [19]. Here, decision in clustering process is made for the number of clusters with $F_{ratio} > F_{statistics}$. Based on the result in Table 7 proposing 9 clusters is reasonable since the value of $F_{ratio} = 2.296 > F_{(8,40)} = 2.18$ at 95% confidence. Where,

$$F_{ratio} = \frac{\text{Between-cluster variability}}{\text{Within-cluster variability}}.$$

Roughly, the number of clusters can be suggested by looking the bend point on the scree plot of within-cluster sum of square (the change in within-groups sum of square error below this point should be negligible)[18–21]. The result described at Fig. 8 suggests nine clusters. The two methods are agree on nine numbers of clusters.

(c) *Bootstrap re-sampling of Ward's method*: Which give statistically significant number of clusters for the desired level of confidence. If the proportion of number of times items are assigned together is at least a desired level of confidence times, then this group is considered as one cluster with the desired level of confidence. E.g. If some groups of items are assigned together, with proportion of number of times grater than or equal to 0.95, thus, these groups of items are considered as one cluster in 95% confidence level). Figure 11 reveals that the number of cluster is 6 at 95% confi-

Table 7 Summary table for within-cluster variability, between-cluster variability and F_{ratio}

No. of clusters	2.00	3.00	4.00	5.00	6.00	7.00	8.0	9.00	10.00	11.00	12.00	13.00	14.00
Belween-cluster variability	1219.46	1063.91	948.85	811.08	691.45	658.29	565.32	563.71	438.92	374.95	348.20	320.78	330.61
Within-cluster viability	550.88	661.89	836.55	958.72	1074.51	1112.05	1252.35	1294.21	1351.96	1366.89	1426.89	1447.97	1461.04
F_ratio	0.45	0.62	0.88	1.18	1.55	1.69	2.22	2.30	3.08	3.65	4.10	4.51	4.42

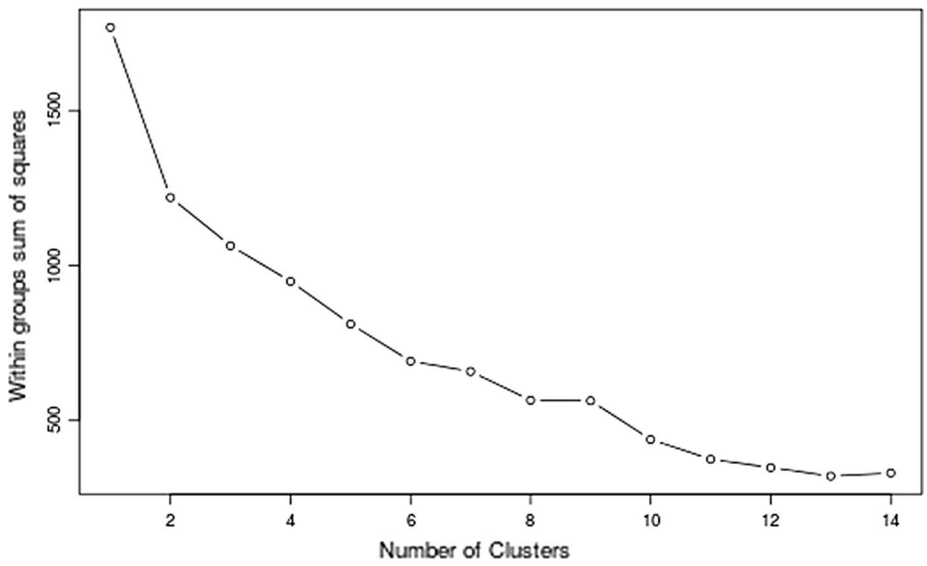

Fig. 8 A scree plot of socio-economic principal factors based on within group sum of squares

dence level and where 5 countries have no data support to be clustered. We should recall that from the detection of outliers by T-Chart I showed those country (South Africa, South Sudan, Niger and Equatorial Guinea) in Fig. 6 plot 8 and in Table 6 as extreme value (outliers) and suspected outliers, except Seychelles. From the above three approach results, I can conclude that, the appropriate number of clusters is 9 considering that some outlier values form individual clusters. They are South Africa, South Sudan and Niger, which each form a cluster.

Determining elements of the cluster

The result of cluster analysis from average linkage method, K-mean method, Ward's method and Bootstrap Wards method were compared. The rough view of the result described in Table 8 suggests that almost all three methods agree on cluster 1, 4, 7, 8. It is an indication for stability of clusters. However, some deviations are observed. For instance K-mean method did not split out South Africa as average linkage and Ward's method do. Average linkage method grouped most of countries in cluster 3 of K-mean method in to other clusters and Wards method merged cluster 3 and 9 of K-mean method in to one cluster (cluster 3). More details of the result for each method are discussed in (a), (b) and (c).

(a) *Average linkage method*: In this method the nearest clusters are joined based on average distance between them, where the distance is the euclidean distance between all items of pairs of clusters. Based on the result of analysis described in Table 8 the deviation of this method is that, most of the countries are grouped in cluster 3 where as by K-mean method these countries are split into three clusters specifically cluster 3, 5 and 9, and into two clusters by Ward's method specifically cluster 3 and 5.

(b) *K-mean method:* It is one of the Non-hierarchical cluster analysis with a purpose of assigning elements to pre-determined clusters, in a way that each item is assigned to a cluster with the nearest mean for the first two principal factors (in this case these two components explained 42.36% of the total variation), while the distance is meas-

Table 8 Summary for cluster element of each clustering Method

Clusters	Countries assigned for each cluster			
	Average linkage	**K-mean**	**Ward's**	**Boot strap Ward's (95% confidence level)**
Cluster 1	NAM, BWA, SWZ, LSO	BWA, NAM, SWZ	NAM, BWA, SWZ, LSO, MUS, DJI	NAM, BWA, SWZ, LSO, MUS, DJI
Cluster 2	DZA, MAR, MUS, TUN, CPV	MUS, TUN, CPV, STP, DJI, LSO, LBR	DZA, MAR, TUN, CPV	DZA, MAR, TUN, CPV
Cluster 3	AGO, MRT, SDN, GHA, ZMB. CMR, STP, DJI, KEN, SEN, BEN, ZWE, TCD, ERI, MLI MDG, UGA, COM, BFA, GNB, SLE, TGO, GIN, RWA, CAF, MOZ, LBR, BDI, MWI, SOM	MRT, CMR, BEN, TCD, MLI, MDG, GIN, MOZ, SOM	TCD, MLI, SOM, MRT, GIN, MOZ, SDN, MDG, GHA, CMR, KEN, SEN, BEN, AGO, ZMB, ZWE	TCD, MLI, SOM, MRT, GIN, MOZ, SDN, MDG, GHA, CMR, KEN, SEN, BEN, AGO, ZMB, ZWE
Cluster 4	NGA, ETH	NGA, SDN, ETH	NGA, ETH	NGA, ETH
Cluster 5	LBY, GAB	GHA, ERI, UGA, COM, BFA, GNB, SLE, TGO, RWA, CAF, BDI, MWI	UGA, MWI, ERI, COM, BFA, GNB, SLE, TGO, RWA, CAF, BDI, STP, LBR	UGA, MWI, ERI, COM, BFA, GNB, SLE, TGO, RWA, CAF, BDI, STP, LBR
Cluster 6	GNQ, SYC	GNQ, LBY, SYC, GAB, DZA, ZAF, MAR	GNQ, LBY, SYC, GAB	LBY, GAB
Cluster 7	SSD	SSD	SSD	
Cluster 8	NER	NER	NER	
Cluster 9	ZAF	AGO, ZMB, KEN, SEN, ZWE	ZAF	

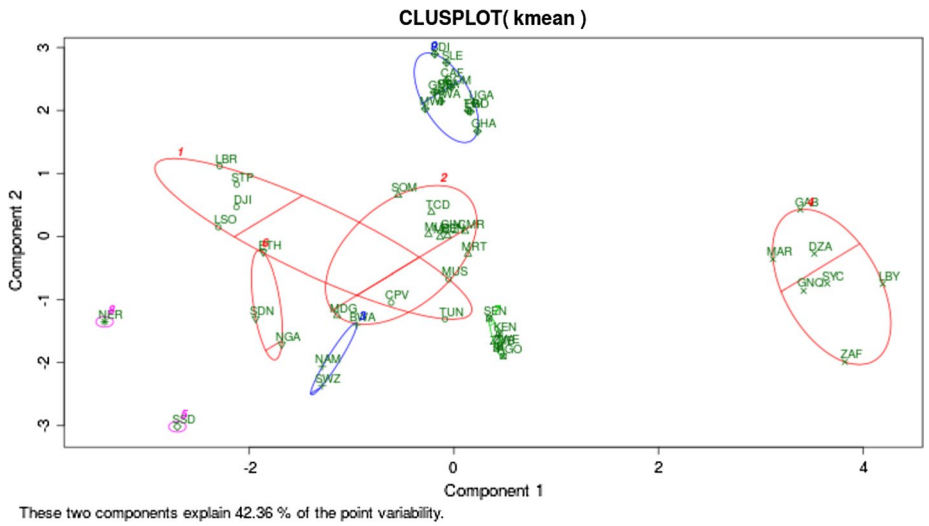

Fig. 9 Clustering by K-mean method of African countries based on the first two socio-economic principal factors

ured by euclidean distance. Based on the result described by Fig. 9 and Table 8, clustering by this method is almost agree with Ward's method, with some exceptions, like, this method merges South Africa but Wards split it out as one cluster, cluster 3 and 9 elements of this method are merged into cluster 3 by Ward's method.

(c) *Ward's method and Bootstrap re-sampling of Ward's method:* The objective of Ward's method is to minimize the information lost in clustering by joining clusters resulting in minimum error sum of square. As described in (b) Ward's method almost agree with K- Mean method with some exceptions. This is an indication of stability of clusters. Additionally, the result from Bootstrap as described in Table 8 assures the existence of the first 6 clusters and stability of its elements at 95% confidence level. Bootstrap suggests that the cluster with large p-value is highly supported by the data. Hence, the number of clusters and elements selection have to be done based on the desired p-value. Based on the result in Fig. 11 there is no enough evidence to reject non-existence of the clusters formed by South Africa, South Sudan, Niger at 95% confidence level. Equatorial Guinea and Seychelles are not included in cluster 6 at this level of confidence (their p-value is 77%), this implies that both Equatorial Guinea and Seychelles are included 77% of times in cluster 6. So including them in cluster 6 will not have big influence on similarity of cluster elements. The Dendrogram representation of Ward's method and Bootstrap re-sampling of Ward's method is given in Figs. 10 and 11 respectively.

Conclusion for cluster analysis

Based on the results of the above four methods discussed, Wald's method finds out the most stable and statistically significantly exist clusters with the exception of cluster 7, 8 and 9. This result strengthen suggestion given on data quality "Data quality" section the 3 left clusters made by South Africa, Niger and South Sudan needs further investigation from the source of data. Meanwhile, since the data source is not easily accessed for further assessment and AfDB [10] was accepted the IMF report, the better decision is taking the three clusters to

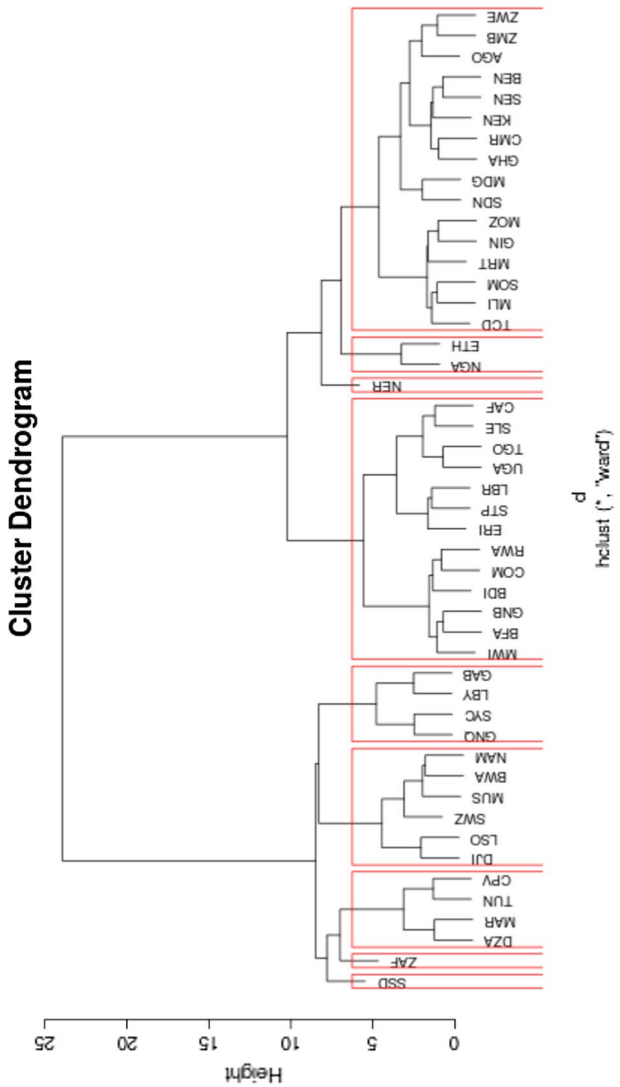

Fig. 10 Clustering by Ward's method of African countries based on socio-economic principal factors

measure their current status. However, further inferences for the population were made based on those statistically significant clusters by including South Africa, because based on real situation observed relatively some extreme data values for South Africa is expected.

Inference for population

The summary result in Table 9, Appendix 1: Tables and Figs. 12, 13 for the relation between clusters and principal factors suggests that cluster 2, 9, 1 and 6 countries have good sustainable (Good) life (variables of PC1, Appendix 1) than other cluster countries. This result specifiably indicates that Tunisia, Mauritius, Seychelles, Cape-Verde, Morocco, Algeria and South Africa have relatively better sustainable life than other African countries, this implies that these countries used relatively suitable policies on variables of PC1 than other African countries used. In terms of capital (variables of PC2, Appendix 1) cluster 9, 2 and 4 countries have a better status. This result specifiably shows that South Africa, Nigeria, Algeria and Morocco have relatively better capital than

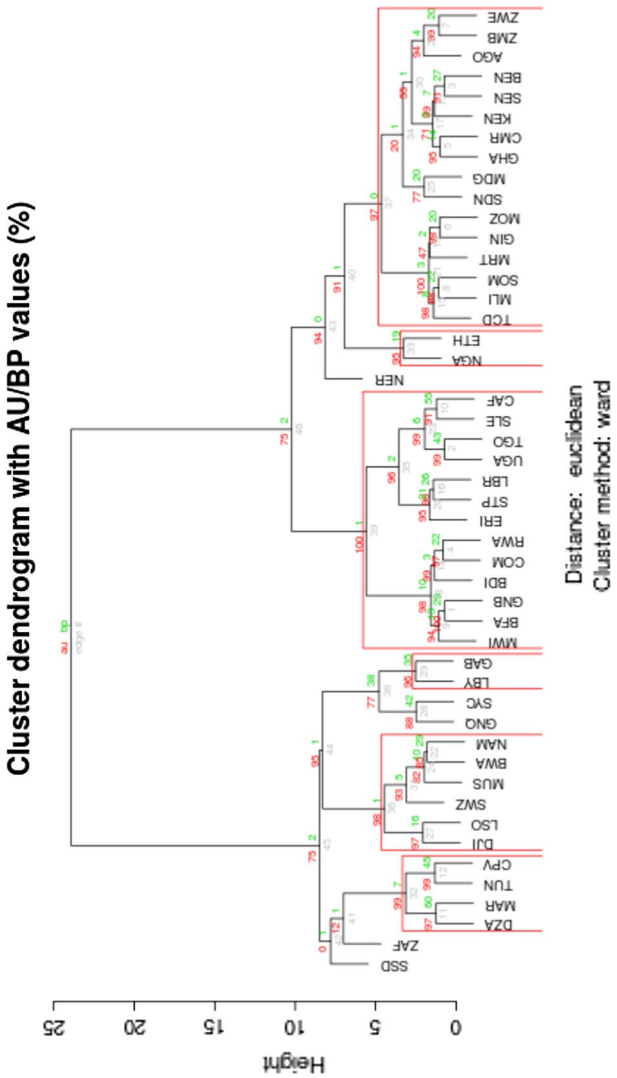

Fig. 11 Clustering by Bootstrap Ward's method of African countries based on socio-economic principal factors

other African countries, this implies that these countries used relatively suitable policies on variables of PC2 than other African countries used.

Cluster 6 countries are generating high income and income related variables (variables of PC3, Appendix 1). So all cluster 6 countries Libya, equatorial Guinea, Seychelles and Gabon policies on PC3 variables are relatively preferable. Life risk (variables of PC4, Appendix 1) is low in cluster 4 and 8 countries, so following Ethiopian, and Nigerian policy in this aspect (for variables of PC4) can reduce life risk (here generalization based on Niger status is on given even if it scores medium PC4 value, since it's cluster is not statistically significant). Cluster 4, and 5 have good literacy (variables of PC5, Appendix 1) status. Mainly Ethiopia, Burundi and Burkina Faso are doing appreciable work on addressing illiteracy reduction. Djibouti, Seychelles, Lesotho and Liberia have relatively better Water supply for domestic consumption contrast to supply for Agriculture (variables of PC6, Appendix 1). Surprisingly, most of cluster countries have no good supply

Table 9 Summary for level of clusters based on principal factors

Ward's cluster	PCI	PC2	PC3	PC4	PC6	PC7	PC5
Cluster 1	Medium (− 0.045, 2.11)	Low (− 0.75, − 0.39)	Low (− 0.75, 1.05)	High (0.59, 3.07)	High (− 0.56, 2.49)	Low (− 0.54, 0.49)	Low (− 0.40, 1.04)
Cluster 2	High level (1.35, 2.46)	Medium (− 0.16,1.44)	Medium (− 0.11, 1.10)	Medium (− 0.54, 0.48)	Low (− 0.71, − 0.05)	Low (− 0.37, 1.12)	Low (− 0.32, 0.57)
Cluster 3	Low (− 1.63, − 0.53)	Low (− 0.69, 0.84)	Low (− 0.96, 0.49)	Medium (− 1.22, 1.24)	Low (− 1.73, − 0.077)	Low (− 0.51, 1.56)	Low (0.65, − 1.13)
Cluster 4	Low (− 1.28, − 0.79)	Medium (0.81, 2.71)	Low (− 0.58, 0.12)	Low (− 1.36, − 1.21)	Low (− 0.93, − 0.72)	Low (− 0.89, 0.41)	High (0.59, 2.71)
Cluster 5	Low (− 1.62, 0.71)	Low (− 0.86, − 0.12)	Low (− 0.41, − 0.41)	Medium (− 0.93, 0.32)	Low (− 0.69, 1.96)	Low (− 1.54 ,0.45)	Medium (− 0.62, 1.21)
Cluster 6	Medium (− 0.13, 2.06)	Low (− 0.53, 0.44)	High (2.23, 3.45)	Medium (− 0.79, 0.61)	Medium (− 1.14, 2.29)	Low (− 0.81, 1.26)	Medium (− 0.13, 0.83)
Cluster 7	Low (− 0.54)	Low (0.47)	Low (0.16)	Medium (− 0.37)	Low (0.35)	High (5.13)	Low (0.03)
Cluster 8	Low (− 0.88)	Low (− 0.34)	Low (− 0.79)	Medium (− 0.95)	Low (− 0.20)	Low (− 0.93)	Low (− 5.03)
Cluster 9	Medium (1.34)	High (5.12)	Medium (0.49)	Medium (2.22)	Low (0.29)	Low (0.17)	Low (− 0.43)

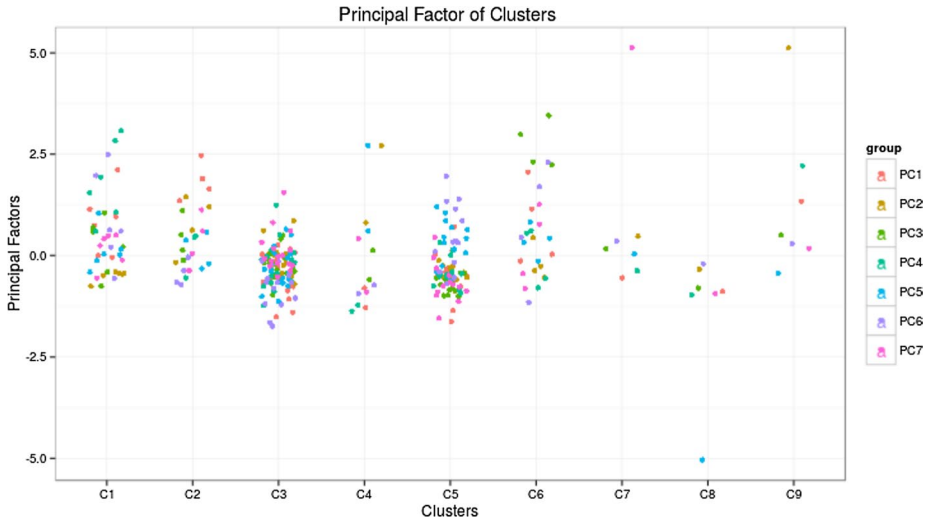

Fig. 12 Dot plot for principal factor values of the clusters

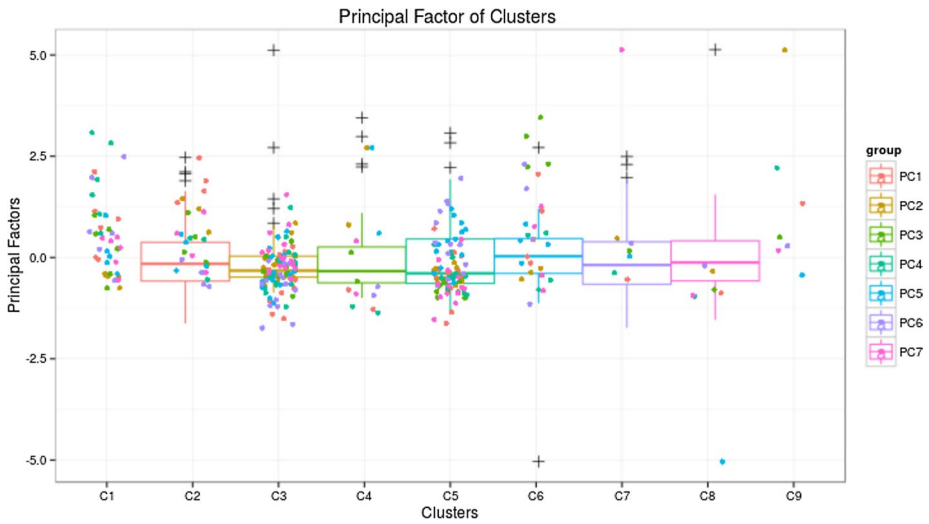

Fig. 13 Dot plot for principal factor values of the clusters and box plot for principal factors

of water, hence addressing pure water for domestic consumption need to be future work of African countries. Angola, Equatorial Genie and Cape Verde have good economic growth (here generalization based on South Sudan status is on given even if it scores the highest PC7 value, since it's cluster is not statistically significant), but the inflation is high (variables of PC7, Appendix 1), so attention need to give to reduce inflation rate.

Future work

Previous study on socio-economic status measurement construct measuring components or variables based on theoretical view of socio-economic stand of an individual or community [1, 2]. However, socio-economic status measurement is still ongoing problem. So to put a hand in solving this problem statistical approach is used to construct measuring components by investigating a natural correlation exist between possible suggested variables, those can able to cluster countries based on their socio-economic status and

level the status by component. Limitation of this study is a comprehensive single measure for a status is not constructed, rather levelling is component-wise and specific suggestion based on the stand of cluster countries for a component variables is given.

Conclusion

The result of Principal component analysis, factor analysis and cluster analysis reveals that, 70% of the variation is encountered by 7 principal factors (Appendix 1: Table 11), using this variation, countries are grouped in to 6 statistically significant (at 95% Confidence Interval) and stable clusters with additional three outlier clusters Table 9. Facts observed from the final out put in Fig. 13 (where the black cross shows outlier values of principal factors) and in Appendix 1: Tables suggests that, Tunisia, Mauritius and Seychelles have relatively better sustainable life (specifically on PC1 Variables listed on Appendix 1: Tables, where as South Africa and Nigeria (recently boom capital) accounts for huge Capital in Africa (specifically for PC2 Variables listed on Appendix 1: Tables. In addition the result also indicates that, Libya, Equatorial Guinea, Seychelles and Gabon have better income source (Mainly income related factors or in general on PC3 Variables listed on Appendix 1: Tables, however, except Seychelles those countries main source of income is oil. Further to these, Ethiopia, Sudan and Nigeria have low life risk or good health policy (specifically on PC4 Variables listed on Appendix 1: Tables. From the result there is also an evidence for a better performance by Ethiopia, Burundi, Burkina Faso and Botswana on literacy reduction (specifically PC5 Variables listed on Appendix 1: Tables. However, it is claimed that, water supply for domestic consumption is not in good status over the continent, even though, Djibouti, Seychelles, Lesotho and Liberia give a better focus for domestic water consumption contrast for Agricultural purpose (generally for PC6 Variables listed on Appendix 1: Tables. It is also pointed that, Angola, Equatorial Guinea and Cape Verde have a better GDP per capita growth, but with high inflation rate (specifically on PC7 variables listed on Appendix 1: Tables. However, high inflation is tackle for growth rate [17], so it needs a solution.

The general suggestion can be, Tunisia's sustainable life policies (variables of PC1, Appendix 1: Tables, South Africa's and Nigeria's Strategy on building Economic Capital, Seychelles's income source policy (oil independent economy), Ethiopia's health and illiteracy reduction policies, Djibouti water supply policy for domestic consumption and Angola's economy growth strategy with some intervention policies on controlling inflation for one country can help to have a better socio-economic status. Specifically, manufacturing areas are comparatively exposed to HIV and tuberculosis, so controlling mechanism should be applied to reduce prevalence rate. In another side, poor sanitation and communicable diseases have correlation with life expectancy and infant mortality rate. Hence, improving sanitation and controlling communicable disease can bring good life expectancy and reduce infant mortality. It is also observed that economic status of a country mainly GDP at market price (current US$) is affected by Foreign direct investment net inflows (BoP, current US$), Gross capital formation (current US$) and transportation system. Hence, adapting economic policy that can attract Foreign direct investment and developing good saving culture with a better transportation system can help to enhance GDP of a county (This result agree with Barro [17] and Upreti [4] suggestion). In addition producing a system which create and use high Electric power and produce high quantities of export of goods and services can help to enhance GDP per capita, PPP (current international $).

Acknowledgements
The author forwards his heartfelt gratitude to two anonymous reviewers for their careful reading of the manuscript and their helpful comments that improve the presentation of this work. Moreover, the author is also grateful for to Prof. Dr. Axel Schumann for his valuable comments. The author also thanks International Monitory fund for free data source and AIMS-Cameroon for resources in doing a paper.

Competing interests
The author declares that he has no competing interests.

Availability of supporting data
All support data files are available.

Consent for publication
Author proves consent of publication for this research.

Appendix 1

See Tables 10, 11, 12, 13, 14, 15 and 16.

Table 10 4 principal factors loadings, variance accounted by factors, and correlation between factors

Variables	PC1	PC2	PC3	PC4	h2	u2	com
NV.AGR.TOTL.ZS	-0.533	-0.150	-0.427	-0.163	0.7557	0.244	2.31
NY.GNS.ICTR.ZS	0.167	-0.115	0.250	0.028	0.1070	0.893	2.24
GC.BAL.CASH.GD.ZS	0.138	0.075	0.144	-0.013	0.0679	0.932	2.52
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
IS.RRS.TOTL.KM	-0.083	0.827	-0.015	0.238	0.6712	0.329	1.19
EG.USE.ELEC.KH.PC	0.435	0.457	0.248	0.120	0.7195	0.280	2.69
IT.NET.BBND.P2	0.402	-0.040	0.193	0.211	0.3076	0.692	2.04

Variance Accounted by factors

	PC1	PC2	PC3	PC4
SS loadings	12.63	7.47	5.98	4.02
Proportion Var	0.23	0.13	0.11	0.07
Cumulative Var	0.23	0.36	0.47	0.54

With component correlations of Principal Factors

	PC1	PC2	PC3	PC4
PC1	1.00	0.30	0.29	0.17
PC2	0.30	1.00	0.25	-0.07
PC3	0.29	0.25	1.00	0.07
PC4	0.17	-0.07	0.07	1.00

Table 11 7 principal components loadings, variance accounted by factors, and correlation between factors

Variables	PC1	PC2	PC3	PC4	PC6	PC7	PC5	h2	u2	com
NV.AGR.TOTL.ZS	-0.406	-0.026	-0.381	-0.319	0.000	-0.165	0.040	0.818	0.1821	3.28
NY.GNS.ICTR.ZS	0.237	0.010	-0.051	-0.117	0.387	0.057	0.060	0.199	0.8007	2.03
GC.BAL.CASH.GD.ZS	-0.024	-0.037	-0.011	0.200	-0.045	0.185	0.765	0.613	0.3866	1.28
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
IS.RRS.TOTL.KM	-0.021	0.893	-0.015	0.135	0.258	-0.063	-0.147	0.789	0.2108	1.29
EG.USE.ELEC.KH.PC	0.184	0.451	0.580	0.150	0.101	-0.163	-0.103	0.826	0.1737	2.64
IT.NET.BBND.P2	0.367	-0.109	0.172	0.246	-0.051	-0.004	-0.078	0.345	0.6546	2.63

Variance Accounted by factors

	PC1	PC2	PC3	PC4	PC6	PC7	PC5
SS loadings	11.48	6.46	5.56	5.28	4.16	3.62	2.89
Proportion Var	0.21	0.12	0.10	0.09	0.07	0.06	0.05
Cumulative Var	0.21	0.32	0.42	0.51	0.59	0.65	0.70

With component correlations of Principal Factors

	PC1	PC2	PC3	PC4	PC6	PC7	PC5
PC1	1.00	0.21	0.45	0.45	0.18	0.17	0.11
PC2	0.21	1.00	0.16	0.06	-0.22	0.14	0.09
PC3	0.45	0.16	1.00	0.19	0.12	0.24	0.10
PC4	0.45	0.06	0.19	1.00	0.35	0.23	-0.05
PC6	0.18	-0.22	0.12	0.35	1.00	0.11	0.01
PC7	0.17	0.14	0.24	0.23	0.11	1.00	-0.09
PC5	0.11	0.09	0.10	-0.05	0.01	-0.09	1.00

Table 12 The result for correlation between key variables in each factor

First.variable	Second.variable	Correlation
Correlation between key variables of principal factor 1		
SP.DYN.CBRT.IN	SP.P0P.1564.TO.ZS	− 0.9849221
SP.DYN.CDRT.IN	SP.DYH.LE00.IN	− 0.9555538
SH.DTH.COMM.Z5	SH.DTH.NCOM.ZS	− 0.8621344
SP.DYH.IMRT.IN	SP.DYN.CDRT.IN	0.8453410
SP.DVN.IMRT.IN	SP.DYN.LE00.IN	− 0.8443482
SP.DYH.IMRT.IN	SH.STA.MMRT	0.8271627
SP.DYN.IHRT.IN	SH.DTH.NCOM.ZS	− 0.8236568
SH.TBS.INCD	SH.DYN.AIDS.ZS	0.8122523
SH.STA.ACSN	SP.P0P.1564.TO.ZS	0.8022663
SP.DYN.IMRT.IN	SP.DVN.CBRT.IN	0.7990561
Correlation between key variables of principal factor 2		
SL.TLF.TOTL.IN	SP.POP.TOTL	0.9320284
NY.GDP.MKTP.CD	NE.GDI.TOTL.CD	0.9250196
NY.GDP.MKTP.CD	BX.KLT.DINV.CD.WD	0.8533420
IS.AIR.DPRT	IS.SHP.GOOD.TU	0.8509953
IS.AIR.DPRT	IS.RRS.TOTL.KM	0.8310989
IS. SHP. GOOD. TU	IS.RRS.TOTL.KM	0.8007450
HY.GDP.MKTP.CD	IS.AIR.DPRT	0.7924445
NE.GDI.TOTL.CD	IS.AIR.DPRT	0.7815439
HY.GDP.MKTP.CD	IS.RRS.TOTL.KM	0.7418154
BX.KLT.DINV.CO.WD	NE.CDI.TOTL.CD	0.7240664
Correlation between key variables of principal factor 3		
NY.GDP.PCAP.CD	NY.GDP.PCAP.PP.CD	0.9849641
NE.EXP.GNFS.ZS	NE.TRD.GNFS.ZS	0.8676546
GC.DOD.TOTL.GD.ZS	NY.GDP.PCAP.CD	0.8034539
GC.DOD.TOTL.GD.ZS	NY.GDP.PCAP.PP.CO	0.7763749
EG.USE.ELEC.KH.PC	NY.GOP.PC AP.CO	0.7654234
EC.USE.ELEC.KH.PC	IT.CEL.SETS.P2	0.7615964
GC.DOD.TOTL.GD.ZS	IT.CEL.SETS.P2	0.7585655
EG.USE.ELEC.KH.PC	NY.GDP.PCAP.PP.CO	0.7530654
NY.GDP.PCAP.PP.CD	SH.STA.ACSN	0.7441146
NY.GDP.PCAP.CD	SH.STA.ACSN	0.7337638
Correlation between key variables of principal factor 4		
SP.DYN.CDRT.IN	SP.DYN.LE00.IN	− 0.955553797
SH.TBS.INCD	SH.DYN.AIDS.ZS	0.812252329
SH.TBS.INCD	SP.DYN.LE00.IN	− 0.497366461
SP.DYN.CDRT.IN	SH.TBS.INCD	0.456132998
NV.IND.MANF.ZS	SH.DYN.AIDS.ZS	0.442521049
SP.DYN.LE00.IN	SH.DYN.AHSS.ZS	0.415258685
SP.DYN.CDRT.IN	SH.DYN.AIDS.ZS	0.368222631
SH.TBS.INCD	W.IND.HANF.ZS	0.362811409
SP.DYN.CDRT.IN	W.IND.rtftNF.ZS	0.012273002
SP.DYN.LE00.IN	W.IND.WAHF.ZS	− 0.008553784
Correlation between key variables of principal factor 5		
SE.ADT.LITR.ZS	SE.ADT.1524.LT.ZS	0.96188933
GC.BAL.CASH.GD.ZS	SE.ADT.1524.LT.ZS	0.72529104
SE.ADT.LITR.ZS	GC.BAL.CASH.GD.ZS	0.70433721
SE.ADT.LITR.ZS	SH.HED.BEDS.ZS	0.38083494

Table 12 continued

First.variable	Second.variable	Correlation
SE.ADT.1524.LT.ZS	SH.MED.BEDS.ZS	0.33375534
GC.BAL.CASH.GD.ZS	SH.MED.BEDS.ZS	0.06028385
SE.ADT.LITR.ZS	SE.ADT.LITR.ZS	0.00000000
GC.BAL.CASH.GD.ZS	SE.ADT.LITR.ZS	0.00000000
SE.ADT.1524.LT.ZS	SE.ADT.LITR.ZS	0.00000000
SH.MED.BEDS.ZS	SE.ADT.LITR.ZS	0.00000000
Correlation between key variables of principal factor 6		
ER.H2O.FWAG.ZS	ER.H2O.FWDM.ZS	− 0.92178608
ER.H2O.FWAG.ZS	ER.H2O.FWIN.ZS	− 0.79363828
ER.H2O.FWDM.ZS	ER.H2O.FWIN.ZS	0.49823355
ER.H2O.FWAG.ZS	ER.H2O.FWTL.K3	0. 38491556
ER.H2O.FWDM.ZS	ER.H2O.FWTL.K3	− 0.37689732
ER.H2O.FWIN.ZS	ER.H2O.FWTL.K3	− 0.26838579
ER.H2O.FWDM.ZS	NV.IND.MANF.ZS	− 0.26534460
ER.H2O.FWAG.ZS	NV.IND.MANF.ZS	0.22052512
ER.H2O.FWTL.K3	NV.IND.MANF.ZS	0.10674733
ER.H2O.FWIN.ZS	NV.IND.MANF.ZS	− 0.09405873
Correlation between key variables of principal factor 7		
NY.GDP.PCAP.KD.ZG	FP.CPI.TOTL.ZG	0.9537706
NE.EXP.GNFS.ZS	NY.GDP.PCAP.KD.ZG	0.4002217
NE.EXP.GNFS.ZS	FP.CPI.TOTL.ZG	0.3793699
NY.GDP.PCAP.KD.ZG	SH.XPD.TOTL.ZS	− 0.3397892
NE.EXP.GNFS.ZS	SH.XPD.TOTL.ZS	− 0.3335359
NE.EXP.GNFS.ZS	SH.HED.BEDS.ZS	0.3101050
SH.XPD.TOTL.ZS	FP.CPI.TOTL.ZG	− 0.2924540
SH.MED.CMHW.P3	NE.EXP.GNFS.ZS	0. 2434526
SH.XPD.TOTL.ZS	SH.MED.PHYS.ZS	− 0.2175899
SH.XPD.TOTL.P3	SH.XPD.TOTL.ZS	− 0.1857653

Table 13 Summary for Ward cluster principal factors

Counlry_name	Country_ code	Ward	PCI	PC2	PC3	PC4	PC6	PC7	PC5
Botswana	BWA	C1	0.9590110484	− 0.3945917923	0.5887618673	1.5471241721	0.633116552	− 0.5498087905	1.0440660176
Djibouti	DJI	C1	0.745605057	− 0.4356574743	− 0.7535622621	1.0676768238	2.4938077094	0.4113095762	− 0.1248642334
Lesotho	LSO	C1	− 0.0456025812	− 0.7564365349	− 0.4015035544	2.8329398055	1.9737241609	0.4999362558	0.0242530618
Mauritius	MUS	C1	2.1152154953	− 0.431877362	1.0553467508	0.5988078203	0.5969491278	− 0.1067921935	0.1548277827
Namibia	NAM	C1	1.147009613	− 0.4563547432	0.2150171211	1.9341183318	0.208735178	0.2448185111	− 0.4078007649
Swaziland	SWZ	C1	0.0005717411	− 0.4811094418	0.6914780197	3.0759689294	− 0.5621275158	0.484863132	0.0309806385
Cabo Verde	CPV	C2	1.8948199088	− 0.1684632366	− 0.1171025924	0.4487535993	− 0.363254217	1.1257812495	− 0.202975648
Algeria	DZA	C2	1.3520498856	1.4480896503	1.1025724821	− 0.5434261143	− 0.6583047633	− 0.370461135	0.5769191795
Morocco	MAR	C2	1.6429869446	1.2089126319	0.1303753392	− 0.0401952201	− 0.7168104136	0.050320077	0.3771171183
Tunisia	TUN	C2	2.4675698087	0.6229568445	0.5139631389	0.4838127146	− 0.0510448912	0.5992877769	− 0.3236315395
Angola	AGO	C3	− 0.759154627	0.0497537961	0.4017519044	0.0637597847	− 0.228800123	1.5583684247	0.0295193365
Benin	BEN	C3	− 0.1670446404	− 0.4667218467	− 0.4266450601	− 0.1769026669	− 0.1388025007	− 0.082260734	− 0.4149003568
Cameroon	CMR	C3	− 0.4807298802	− 0.1760008009	− 0.218343622	− 0.1026342918	− 0.6501814707	− 0.3696230457	0.2227357734
Ghana	GHA	C3	0.150675033	0.0280875925	− 0.1893003346	− 0.4274158494	− 0.1544634325	− 0.2612943406	0.6508580273
Guinea	GIN	C3	− 0.6471937337	− 0.4316969358	− 0.063392055	− 0.6639929574	− 0.0769017334	− 0.1824883993	− 1.0064768849
Kenya	KEN	C3	− 0.0298161108	0.6164543168	− 0.5266114576	0.1229507539	− 0.5644707004	0.3318258651	− 0.061297445
Madagascar	MDG	C3	0.0770485667	− 0.0061854139	− 0.6673556985	− 0.4978087997	− 1.735873743	0.8121329242	− 0.7494811779
Mali	MLI	C3	− 0.8650614931	− 0.3755346296	− 0.3836473499	− 0.6521257498	− 1.2027659092	− 0.2484179546	− 1.1292854938
Mozambique	MOZ	C3	− 1.0631387581	0.0619231978	− 0.3222168552	− 0.0965777327	− 0.4210766747	− 0.1993556418	− 0.6699729633
Mauritania	MRT	C3	− 0.5498164081	− 0.6969453485	0.4944055403	− 0.6310346354	− 1.1855611494	0.176677758	− 0.6444569793
Sudan	SDN	C3	− 0.1431538132	0.8491909175	− 0.3488061831	− 1.2203951774	− 1.6489727741	− 0.1411503783	0.5087899484
Senegal	SEN	C3	0.2650745701	− 0.2512365858	− 0.477922001	− 0.0118605921	− 0.6605989007	− 0.0052064657	− 0.3691277625
Somalia	SOM	C3	− 1.3947612484	− 0.6089330356	− 0.9622126227	− 0.876429411	− 1.0441987588	− 0.5182244058	− 0.5496406506
Chad	TCD	C3	− 1.5086120175	− 0.4952357188	− 0.2344523805	− 0.7581924139	− 0.66051074	− 0.0133297006	0.0587157495
Zambia	ZMB	C3	− 0.3011832782	0.0214364393	0.0328310111	0.5124679935	− 0.4813970297	0.6151853365	− 0.5802166966
Zimbabwe	ZWE	C3	0.0306690278	− 0.2475545317	− 0.0997898085	1.2370515845	− 0.7966948311	0.12535114	− 0.337998809

Table 13 continued

Counlry_name	Country_ code	Ward	PCI	PC2	PC3	PC4	PC6	PC7	PC5
Ethiopia	ETH	C4	− 0.7939023513	0.8092267712	− 0.5824325895	− 1.3668481129	− 0.9362386016	− 0.8943634801	2.7142917656
Nigeria	NGA	C4	− 1.2831650235	2.7135554824	0.1218338427	− 1.2128350948	− 0.7188175904	0.4124323917	0.5970206693
Burundi	BDI	C5	− 0.5578081702	− 0.4609954697	− 0.9974226677	− 0.7465524246	0.094834108	− 1.5383724936	1.2063159848
Burkina Faso	BFA	C5	− 0.3202038684	− 0.4163876658	− 0.8531724603	− 0.4585173205	− 0.6936412091	− 0.8681811158	1.0531878188
Central African Republic	CAF	C5	− 1.3491511679	− 0.2519465068	− 0.6705583842	− 0.4644432333	1.1505640591	− 1.1279203503	− 0.6288008228
Comoros	COM	C5	0.034565117	− 0.6875390374	− 0.5888599785	− 0.520034862	0.3312569286	− 0.8893110221	0.8612547019
Eiitiea	ERI	C5	− 0.372783211	− 0.5939013817	− 0.5459384885	− 0.9274709521	1.3923504161	0.4467761909	0.3124683982
Guinea− Bissau	GNB	C5	− 0.7525642439	− 0.5336574175	− 0.6102687776	− 0.4431306963	− 0.3982839423	− 0.7391923422	0.6920937892
Liberia	LBR	C5	− 0.2910310997	− 0.8667475677	− 0.4241000099	0.1409091409	1.9569699036	− 0.3167840488	− 0.3868496662
Malawi	MWI	C5	− 0.3955179182	− 0.5139131604	− 0.9762428733	0.3211653565	− 0.5402357928	− 0.7005865311	0.4553645932
Rwanda	RWA	C5	0.1228870325	− 0.303957657	− 0.982927951	− 0.505869568	− 0.1673233739	− 0.6531985683	0.6298648896
Sierra Leone	SLE	C5	− 1.6280307567	− 0.2687501892	− 0.8239875566	− 0.8635193488	0.8610836584	− 0.9742680075	0.4240034914
Sao Tome and Principe	STP	C5	0.7135294396	− 0.5007115244	− 0.8354318376	0.0044234779	1.3382960688	− 0.0497183385	0.0587667358
Togo	TGO	C5	− 0.5366400122	− 0.5467933019	− 0.4176095268	− 0.4823314209	0.3212694991	− 0.4803872208	− 0.2569195743
Uganda	UGA	C5	− 0.5098114743	− 0.1244103371	− 0.7061415117	− 0.406426556	0.3680250101	− 0.7644691821	0.1519163605
Gabon	GAB	C6	0.0254783415	− 0.3714064901	2.2385339766	− 0.5526971911	0.4464879495	− 0.8128174113	0.8322568669
Equatorial Guinea	GNQ	C6	− 0.1334869541	− 0.2687146646	2.9876836758	0.5478497756	1.6958489031	1.2575918528	− 0.1349036955
Libya	LBY	C6	1.1520358054	0.4412331329	3.4529067947	− 0.7913939194	− 1.1462088237	− 0.440047964	0.4287180129
Seychelles	SYC	C6	2.0643701037	− 0.5314056878	2.3077120686	0.61025422	2.2977538318	0.7729174579	0.3206752388
South Sudan	SSD	C7	− 0.541103311	0.47262448	0.1677652081	− 0.3714270057	0.3543032925	5.1315033763	0.0310225258
Niger	NER	C8	− 0.881311579	− 0.3421598744	− 0.7935467496	− 0.957194485	− 0.200538362	− 0.9315367558	− 5.0350218898
South Africa	ZAF	C9	1.3414077913	5.1204881128	0.4986170192	2.219569511	0.2887236111	0.172488721	− 0.4333814221

Table 14 Clusters of Country sorted by principal component

Country	Ward	PC1	Country	Ward	PC2	Country	Ward	PC3	Country	Ward	PC4	Country	Ward	PC6	Country	Ward	PC7	Country	Ward	PC5
TUN	2	2.4676	ZAF	9	5.1205	LBY	6	3.4529	SWZ	1	3.076	DJI	1	2.4938	SSD	7	5.1315	ETH	4	2.7143
MUS	1	2.1152	NGA	4	2.7136	GNQ	6	2.9877	LSO	1	2.8329	SYC	6	2.2978	AGO	3	1.5584	BDI	5	1.2063
SYC	6	2.0644	DZA	2	1.4481	SYC	6	2.3077	ZAF	9	2.2196	LSO	1	1.9737	GNQ	6	1.2576	BFA	5	1.0532
CPV	2	1.8946	MAR	2	1.2089	GAB	6	2.2385	NAM	1	1.9341	LBR	5	1.957	CPV	2	1.1258	BWA	1	1.0441
MAR	2	1.643	SDN	3	0.8492	DZA	2	1.1026	BWA	1	1.5471	GNQ	6	1.6958	MDG	3	0.8121	COM	5	0.8613
DZA	2	1.352	ETH	4	0.8092	MUS	1	1.0553	ZWE	3	1.2371	ERI	5	1.3924	SYC	6	0.7729	GAB	6	0.8323
ZAF	9	1.3414	TUN	2	0.623	SWZ	1	0.6915	DJI	1	1.0677	STP	5	1.3383	ZMB	3	0.6152	GNB	5	0.6921
LBY	6	1.152	KEN	3	0.6165	BWA	1	0.5888	SYC	6	0.6103	CAF	5	1.1506	TUN	2	0.5993	GHA	3	0.6509
NAM	1	1.147	SSD	7	0.4726	TUN	2	0.514	MUS	1	0.5988	SLE	5	0.8611	LSO	1	0.4999	RWA	5	0.6299
BWA	1	0.959	LBY	6	0.4412	ZAF	9	0.4986	GNQ	6	0.5478	BWA	1	0.6331	SWZ	1	0.4849	NGA	− 1	0.597
DJI	1	0.7456	MOZ	3	0.0619	MRT	3	0.4944	ZMB	3	0.5125	MUS	1	0.5969	ERI	5	0.4468	DZA	2	0.5769
STP	5	0.7135	AGO	3	0.0498	AGO	3	0.4018	TUN	2	0.4838	GAB	6	0.4465	NGA	4	0.4124	SDN	3	0.5088
SEN	3	0.2651	GHA	3	0.0281	NAM	1	0.215	CPV	2	0.4488	UGA	5	0.368	DJI	1	0.4113	MWI	5	0.4554
GHA	3	0.1507	ZMB	3	0.0214	SSD	7	0.1678	MWI	5	0.3212	SSD	7	0.3543	KEN	3	0.3318	LBY	6	0.4287
RWA	5	0.1229	MDG	3	− 0.0062	MAR	2	0.1304	LBR	5	0.141	COM	5	0.3313	NAM	1	0.2448	SLE	5	0.424
MDG	3	0.077	UGA	5	− 0.1244	NGA	4	0.1218	KEN	3	0.123	TGO	5	0.3213	MRT	3	0.1767	MAR	2	0.3771
COM	5	0.0346	CPV	2	− 0.1685	ZMB	3	0.0328	AGO	3	0.0638	ZAF	9	0.2887	ZAF	9	0.1725	SYC	6	0.3207
ZWE	3	0.0307	CMR	3	0.176	GIN	3	0.0634	STP	5	0.0044	NAM	1	0.2087	ZWE	3	0.1254	ERI	5	0.3125
GAB	6	0.0255	ZWE	3	− 0.2476	ZWE	3	− 0.0998	SEN	3	− 0.0119	BDI	5	0.0948	MAR	2	0.0503	CMR	3	0.2227
SWZ	1	0.0006	SEN	3	− 0.2512	CPV	2	− 0.1171	MAR	2	− 0.0402	TUN	2	− 0.051	SEN	3	− 0.0052	MUS	1	0.1548
KEN	3	− 0.0298	CAF	5	− 0.2519	GHA	3	− 0.1893	MOZ	3	− 0.0966	GIN	3	− 0.0769	TCD	3	− 0.0133	UGA	5	0.1519
LSO	1	− 0.0456	GNQ	6	− 0.2687	CMR	3	− 0.2183	CMR	3	− 0.1026	BEN	3	− 0.1388	STP	5	− 0.0497	STP	5	0.0568
GNQ	6	− 0.1335	SLE	5	− 0.2688	TCD	3	− 0.2345	BEN	3	− 0.1769	GHA	3	− 0.1545	BEN	3	− 0.0823	TCD	3	0.0587
SDN	3	− 0.1432	RWA	5	0.304	MOZ	3	− 0.3222	SSD	7	− 0.3714	RWA	5	− 0.1673	MUS	1	− 0.1068	SSD	7	0.031
BEN	3	− 0.167	NER	8	− 0.3422	SDN	3	− 0.3488	UGA	5	− 0.4064	NER	8	− 0.2005	SDN	3	− 0.1412	SWZ	1	0.031

Table 14 continued

Country	Ward	PC1	Country	Ward	PC2	Country	Ward	PC3	Country	Ward	PC4	Country	Ward	PC6	Country	Ward	PC7	Country	Ward	PC5
LBR	5	− 0.2918	GAB	6	− 0.3714	MLI	3	− 0.3836	GHA	3	− 0.4274	AGO	3	− 0.2288	GIN	3	− 0.1825	AGO	3	0.0295
ZMB	3	− 0.3012	MLI	3	− 0.3755	LSO	1	− 0.4015	GNB	5	− 0.4431	CPV	2	− 0.3633	MOZ	3	− 0.1994	LSO	1	0.0243
BFA	5	− 0.3202	BWA	1	− 0.3946	TGO	5	− 0.4176	BFA	5	− 0.4585	GNB	5	− 0.3983	MLI	3	− 0.2484	KEN	3	− 0.0613
ERI	5	− 0.3728	BFA	5	− 0.4164	LBR	5	− 0.4242	CAF	5	− 0.4644	MOZ	3	− 0.4211	GHA	3	− 0.2613	DJI	1	− 0.1249
MWI	5	− 0.3955	GIN	3	− 0.4317	BEN	3	− 0.4266	TGO	5	− 0.4823	ZMB	3	− 0.4814	LBR	5	− 0.3168	GNQ	6	− 0.1349
CMR	3	− 0.4807	MUS	1	− 0.4319	SEN	3	− 0.4779	MDG	3	− 0.4978	MWI	5	− 0.5402	CMR	3	− 0.3696	CPV	2	− 0.203
UGA	5	− 0.5098	DJI	1	− 0.4357	KEN	3	− 0.5266	RWA	5	− 0.5059	SWZ	1	− 0.5621	DZA	2	− 0.3705	TGO	5	− 0.2569
TGO	5	− 0.5366	NAM	1	− 0.4564	ERI	5	− 0.5459	COM	5	− 0.52	KEN	3	− 0.5645	LBY	6	− 0.44	TUN	2	− 0.3236
SSD	7	− 0.5411	BDI	5	0.461	ETH	4	− 0.5824	DZA	2	− 0.5434	CMR	3	− 0.6502	TGO	5	− 0.4804	ZWE	3	− 0.338
MRT	3	− 0.5498	BEN	3	− 0.4667	COM	5	− 0.5889	GAB	6	− 0.5527	DZA	2	− 0.6583	SOM	3	− 0.5182	SEN	3	− 0.3691
BDI	5	− 0.5578	SWZ	1	− 0.4811	GNB	5	− 0.6103	MRT	3	− 0.631	TCD	3	− 0.6605	BWA	1	− 0.5498	LBR	5	− 0.3868
GIN	3	− 0.6472	TCD	3	− 0.4952	MDG	3	− 0.6674	MLI	3	− 0.6521	SEN	3	− 0.6606	RWA	5	− 0.6532	NAM	1	− 0.4078
GNB	5	− 0.7526	STP	5	− 0.5007	CAF	5	− 0.6706	GIN	3	− 0.664	BFA	5	− 0.6936	MWI	5	− 0.7006	BEN	3	− 0.4149
AGO	3	− 0.7592	MWI	5	− 0.5139	UGA	5	− 0.7061	BDI	5	− 0.7466	MAR	2	− 0.7168	GNB	5	− 0.7392	ZAF	9	− 0.4334
ETH	4	− 0.7939	SYC	6	− 0.5314	DJI	1	− 0.7536	TCD	3	− 0.7582	NGA	4	− 0.7188	UGA	5	− 0.7645	SOM	3	− 0.5496
MLI	3	− 0.8651	GNB	5	− 0.5337	NER	8	− 0.7935	LBY	6	− 0.7914	ZWE	3	− 0.7967	GAB	6	− 0.8128	ZMB	3	− 0.5802
NER	8	− 0.8813	TGO	5	− 0.5468	SLE	5	− 0.824	SLE	5	− 0.8635	ETH	4	− 0.9362	BFA	5	− 0.8682	CAF	5	− 0.6288
MOZ	3	− 1.0631	ERI	5	− 0.5939	STP	5	− 0.8354	SOM	3	− 0.8764	SOM	3	− 1.0442	COM	5	− 0.8893	MRT	3	− 0.6445
NGA	4	− 1.2832	SOM	3	− 0.6089	BFA	5	− 0.8532	ERI	5	− 0.9275	LBY	6	− 1.1462	ETH	4	− 0.8944	MOZ	3	− 0.67
CAF	5	− 1.3492	COM	5	− 0.6875	SOM	3	− 0.9622	NER	8	− 0.9572	MRT	3	− 1.1856	NER	8	− 0.9315	MDG	3	− 0.7495
SOM	3	− 1.3948	MRT	3	− 0.6969	MWI	5	− 0.9762	NGA	4	− 1.2128	MLI	3	− 1.2028	SLE	5	− 0.9743	GIN	3	− 1.0065
TCD	3	− 1.5086	LSO	1	− 0.7564	RWA	5	− 0.9829	SDN	3	− 1.2204	SDN	3	− 1.649	CAF	5	− 1.1279	MLI	3	− 1.1293
SLE	5	− 1.628	LBR	5	− 0.8667	BDI	5	− 0.9974	ETH	4	− 1.3668	MDG	3	− 1.7359	BDI	5	− 1.5384	NER	8	− 5.035

Table 15 Variables of principal components and summary result

	Variable code	Loadinqs	Corr	Com
Principal component 1 (sustainable lite)				
Life expectancy at birth, total (years)	SP.DYN.LE00.IN	1.06	0.8079716	0.9380648
Cause of dealt, by non— commuricable diseases (% of total)	SH.DTH.NCOM.ZS	1.01	0.823052	0.7976794
Improved water source, rural (% of rural population with access)	SH.H2O.SAFE.RU.Z	0.78	0.7575063	0.6749383
Population, ages 15–64 (% of total)	SP.POP.1564.TO.ZS	0.73	0.8764492	0.8405396
Internet users (per 100 people)	IT.NET.USER.P2	0.67	0.7321589	0.6816921
Services, etc., value added (% of GDP)	NV.SRV.TETC.ZS	0.63	0.5728387	0.7281336
Improved sanitation facilities (% of population will access)	SH.STA.ACSN	0.55	0.77288	0.7755206
Improved water source, urban (% ol urban population with access)	SH.H2O.SAFE.URZ	0.54	0.5288412	0.6275978
Gross enrolment raio, pre-primary, both sexes (%)	SE.PRE.ENRR	0.48	0.5518474	0.394378
Mobile cellular subscriptions (per 100 people)	IT.CEL.SETS.P2	0.47	0.7351707	0.8132745
Gross enrolment raio, tertiary, both sexes (%)	SE.TER.ENRR	0.42	0.5128509	0.4215655
Gross enrolment raio, secondary, both sexes (%)	SE.SEC.ENRR	0.39	0.552263	0.4368579
Physicians (per 1000 people)	SH.MED.PHYS.ZS	0.38	0.351458	0.4507551
Fixed broadband subscriptions (per 100 people)	IT.NET.BBNO.P2	0.37	0.51236	0.345433
Prevalence of HIV. total (% of population ages 15–49)	SH.DYN.ADS.ZS	0.41	0.777566	0.8047301
Agriculture, value added (% of GDP)	NV.AGR.TOTL.ZS	0.41	− 0.7452145	0.8178983
Incidence of tuberculosis (per 100,000 people)	SH.TBS.INCD	− 0.53	− 0.0656801	0.7888127
Birth rate, crude (per 1000 people)	SP.DYN.CBRT.IN	− 0.74	− 0.8915574	0.8582985
Maternal mortality ratio (modeled estmate. per 100,000 live births)	SH.STA.MMRT	− 0.82	0.8261328	0.725329
Cause of dealt, by communicable diseases and maternal, prenatal and nutrition conditions (% of total)	SH.DTH.COMM.ZS	− 0.93	− 0.7800013	0.7496144
Mortality rate, infant (per 1000 live births)	SP.DYN.IMRT.IN	0.98	− 0.8936826	0.8558577
Death rate, crude (per 1000 people)	SP.DYN.CDRT.IN	− 1.04	− 0.7554875	0.8471443
Principal component 2 (Captial)				
GDP at market prices (current US$)	NY.GDP.MKTP.CD	0.97	0.9517201	0.9170657

Table 15 continued

	Variable code	Loadinqs	Corr	Com
Rail lines (total roule-km)	IS.RRS.TOTL.KM	0.89	0.8176941	0.7891843
Foreign direct investment, net inflows (BoP. current US$)	BX.KLT.DINV.CD.WD	0.87	0.8659704	0.8049668
Gross capital formation (current US$)	NE.GDI.TOTL.CD	0.85	0.87816171	0.8110496
Air transport, registered carrier departures worldwide	IS.AIR.DPRT	0.84	0.8702821	0.8434615
Container port traffic (TEU: 20 foot equivalent units)	IS.SHP.GOOD.TU	0.74	0.7496073	0.7518197
Population, total	SP.POP.TOTL	0.7	0.6570489	0.6306013
Labor force, total	SL.TLF.TOTL.IN	0.66	0.6468067	0.7005713
Electric power consumption (kWh per capitia)	EG.USE.ELEC.KH.P	0.45	0.5377256	0.8262583
Gross enrolment ratio, tertiary, both sexes (%)	SE.TER.ENRR	0.41	0.4621814	0.4215655
Annual freshwater withdrawals, total (billion cubic meters)	ER.H2O.FWTL.K3	0.34	0.457826	0.4011402
Imports of goods and services (% of GDP)	NE.IMP.GNFS.ZS	− 0.31	− 0.3255368	0.5675932
Principal component 3 (income related factor)				
GDP per capita, PPP (current international $)	NY.GDP.PCAP.PP.C	0.89	0.9446474	0.9213499
GDP per capita (current US$)	NY.GDP.PCAP.CD	0.85	0.9250355J	0.9110524
Industry, value added (% of GDP)	NV.IND.TOTL.ZS	0.83	0.7730768	0.6893409
Central government debt, total (%of GDP)	GC.DOD.TOTL.GD.Z	0.7	0.8077165	0.741978
Exports of goods and services (%of GDP)	NE.EXP.GNFS.ZS	0.59	0.7130032	0.853719
Electric power consumption (kWh per captia)	EG.USE.ELEC.KH.	0.58	0.726317	0.8262583
Mobile cellular subscriptions (per 100 people)	IT.CELSETS.P2	0.55	0.7422919	0.8132745
Improved sanitation facilities (% of population with access)	SH.STA.ACSN	0.46	0.6969131	0.7755206
Trade (%of GDP)	NE.TRD.GNFS.ZS	0.45	0.5497256	0.7607122
Hospital beds (per 1000 people)	SH.MED.BEDS.ZS	0.45	0.4365858	0.4086245
Gross enrolment ratio, secondary, both sexes (9t)	SE.SEC.ENRR	0.35	0.5159998	0.4368579
Improved water source, urban (% of urban population with access)	SH.H2O.SAFE.UR.Z	− 0.31	− 0.0620451	0.6275978
Agriculture, value added (% of GDP)	NV.AGR.TOTL.ZS	0.38	− 0.6669364	0.8178983
Health expenditure, total (% of GDP)	SH.XPD.TOTL.ZS	− 0.46	− 0.47214621	0.5657422
Services.etc., value added (% of GDP)	NV.SRV.TETC.ZS	− 0.5	− 0.0756577	0.7281336

Table 16 Variables of principal components and summary result

	Variable code	Loadings	Corr	Com
Principal component 4 (life risk)				
Prevalences HIV, total (%of population ages 15–49)	SH.DYN.AIDS.ZS	1.05	0.7905701	0.8047301
Incidence of tuberculosis (per 100.000 people)	SH.TBS.INCD	1	0.7393767	0.7888127
Manufacturing, value added (% of G DP)	NV.IND.MANF.ZS	0.77	0.6267752	0.5451881
Death rate, crude (per 1000 people)	SP.DYN.CDRT.IN	0.52	0.1150051	0.8471443
Container port traffic (TEU: 20 fool equivalent unit)	IS.SHP.GOOD.TU	0.39	0.04575752	0.7518197
Imports of goods and Services (% of GDP)	NE.IMP.GNFS.ZS	0.38	0.5374423	0.5675632
Government expenditure on education is % Of GDP (%)	SE.XPD.TOTL.GD.Z	0.38	0.3719803	0.3151677
Trade (% Of GDP)	NE.TRD.GNFS.ZS	0.34	0.5519894	0.7607122
Health expenditure, total (% Of GDP)	SH.XPD.TOTL.ZB	0.33	0.2408635	0.5657422
Improved water source, urban (% of urban population with access)	SH.H2O.SAFE.UR.Z	0.31	0.4626387	0.6275678
Agriculture, value added (% of GDP)	NV.AGR.TOTL.ZS	− 0.32	4.0058561	0.8178983
Total debt service (% of exported goods. servicer and primary income)	DT.TDS.DECT.EX.Z	− 0.37	− 0.1428698	0.2930229
Life expectancy at birth, total (years)	SP.DYN.LE00.IN	− 0.57	− 0.1076823	0.9380648
Principal component 5 (literacy)				
Youth literacy rate, population 15–24 years, both sexes (%)	SE.ADT.1524.LT.ZS	0.87	0.8778255	0.8556102
Adult literacy rate, population 15 + years, both sexes (%)	SE.ADT.LITR.ZS	0.86	0.8747162	0.8454292
Cash surplus/deficit (% of GDP)	GC.BAL.CASH.GD.Z	0.77	0.7345393	0.61338
Hospital beds (per 1000 people)	SH.MED.BEDS.ZS	0.37	0.4346196	0.4086245
Teenage mothers (% of women ages 15–19 who have had children or are currently pregnant)	SP.MTR.1519.ZS	− 0.38	− 0.3940773	0263375
Principal component 6 (rate of water supply and consumption contrast)				
Annual freshwater withdrawals, domestic (% of total freshwater withdrawal)	ER.H2O.FWDM.ZS	1	0.8787253	0.8500172
Annual freshwater withdrawals, industry (% of total freshwater withdrawal)	ER.H2O.FWIN.ZB	0.66	0.5969486	0.478074
Cross savings (% of GDP)	NY.GNS.ICTR.ZS	0.39	0.3865365	0.1993376
Total debt service (% of exports of goods, services and primary income)	DT.TDS.DECT.EX.Z	0.36	0.2825562	0.2930229
Imports of goods and services (% of GDP)	NE.IMP.GNFS.ZS	0.32	0.5611142	0.5675932
Trade (% of GDP)	NE.TRD.GNFS.ZS	0.31	0.5644163	0.7607122
Annual freshwater withdrawals, total (billion cubic meters)	ER.H2O.FWTL.K3	− 0.41	− 0.4826178	0.4011402
Manufacturing , value added (%of GDP)	NV.IND.MANF.ZS	− 0.43	− 0.1210153	0.5451881
Annual freshwater withdrawals, agriculture (% of total freshwater withdrawal)	ER.H2O.FWAG.ZS	− 0.99	− 0.8304037	0.845570E
Principal component 7				
GDP per capita growth (annual %)	NY.GDP.PCAP.KDZ	0.88	0.825833	0.7268232
Inflation, consumer prices (annual %)	FP.CPI.TOTL.ZG	0.86	0.7815264	0.7317425
Community health workers (per 1000 people)	SH.MED.CMHW.P3	0.46	0.4290214	0.2630565
Exports of goods and services (% of GDP)	NE.EXP.GNFS.ZS	0.45	0.6200583	0.853719
Nurses and midwives (per 1000 people)	SH.MED.NUMW.P3	0.4	0.4986513	0.4551595
Physicians (per 1000 people)	SH.MED.PHYS.ZS	0.39	0.4168554	0.4507651
Services, etc., value added (% of GDP)	NV.SRV.TETC.ZS	0.32	0.3575785	0.7281336
Improved water source, urban (% of urban population with access)	SH.H2O.SAFE.UR.Z	− 0.41	− 0.3178607	0.6275978
Health expenditure, total (% Of GDP)	SH.XPD.TOTL.ZS	− 0.46	− 0.4648261	0.5657422

Publisher's Note

Received: 1 August 2017 Accepted: 19 October 2017
Published online: 19 December 2017

References

1. Cowan CD, Hauser RM, Kominski RA, Levin HM, Lucas SR, Morgan SL, Spencer MB, Chapman C. Improving the measurement of socioeconomic status for the national assessment of educational progress: a theoretical foundation. 2012.
2. Oakes JM, Rossi PH. The measurement of SES in health research: current practice and steps toward a new approach. Soc Sci Med. 2003;56(4):769–84.
3. Haller AO. The social grading of occupations: a new approach and scale. In: John H, editors. Goldthorpe Keith Hope; 1976.
4. Upreti P. Factors affecting economic growth in developing Countries. Major themes in economics. Berlin: Spring; 2015.
5. Cochran WG. Sampling techniques. New York: Wiley; 2007.
6. Vaseghi SV. Advanced digital signal processing and noise reduction. New York: Wiley; 2008.
7. Lokupitiya RS, Lokupitiya E, Paustian K. Comparison of missing value imputation methods for crop yield data. Environmetrics. 2006;17(4):339–49.
8. Howell DC. The treatment of missing data. In: The sage handbook of social science methodology; 2007. p. 208–224.
9. Jerven M. Why We Need to Invest in African Development Statistics: From a Diagnosis of Africa'sStatistical Tragedy Towards a Statistical Renaissance. African Arguments 2013.
10. African Development Bank. Situational analysis of economic statistics in Africa: Special focus on GDP measurement. Abidjan: African Development Bank; 2013. http://www.afdb.org/fileadmin/uploads/afdb/Documents/Publications/Economic%20Brief%20-%20Situational%20Analysis%20of%20the%20Reliability%20of%20Economic%20Statistics%20in%20Africa-%20Special%20Focus%20on%20GDP%20Measurement.pdf. Accessed 7 Dec 2017.
11. World Bank. World Economic and Financial Surveys, Regional Economic Outlook, Sub-Saharan Africa. 2008.
12. Miguez F. Introduction to R for multivariate data analysis. 2007.
13. Johnson RA, Wichern DW. Prentice hall Englewood Cliffs. Applied multivariate statistical analysis. 5th ed. New Jersy: Prentice hall Englewood Cliffs; 2002.
14. Rencher AC. Methods of multivariate analysis, vol. 492. New York: Wiley; 2003.
15. Kumar S, Toshniwal D. Analysis of hourly road accident counts using hierarchical clustering and cophenetic correlation coefficient (CPCC). J Big Data. 2016;3(13):1–11.
16. Kumar S, Toshniwal D. A novel framework to analyze road accident time series data. J Big Data. 20016;3(8):1–11.
17. Barro RJ. Determinants of economic growth: a cross-country empirical study (No. w5698). Nat Bureau Econom Res. 1996.
18. Calinski RB, Harabasz J. A dendrite method .for cluster analysis. Commun Stat. 1974;3(1):1–27.
19. Steinley Douglas, Brusco Michael J. Choosing the number of clusters in K-means clustering. Psychol Methods. 2011;3(16):285–97.
20. Steinley D. Validating clusters with the lower bound for sum of squares error. Psychometrika. 2007;72(1):93–106.
21. Steinley D, Brusco MJ. A new variable weighting and selection procedure for K-means cluster analysis. Multivar Behav Res. 2008;43(1):77–108.