ECONOMIC INDICATORS AND HUMAN DEVELOPMENT INDEX

Ana-Maria Mihaela IORDACHE [1*]
Ionela-Cătălina ZAMFIR [2]

ABSTRACT

Human Development Index (HDI) is a very interesting index that may show the development level of a country. A country that has a big HDI index and great life expectancy at birth should be a country with a developed economy, low unemployment rates, good import and export indicators and business-favorable economic environment. The main hypothesis tested in this paper is if there is a relation between HDI index and other type of mentioned indicators. Starting from the idea that in a cybernetic system, there is interdependence between all the events that take place. There are two types of analyses used to reduce the size of 22 variables from the dataset, while K-means and Ward's are used to classify the observations in four classes. A confusion matrix calculated between new classes (K-means algorithm) and known classes (from HDI index) confirm the tested hypothesis, with a certain "accuracy rate".

KEYWORDS: *Dimension reduction analyses, HDI, K-Means, Ward's method, correlation*

JEL CLASSIFICATION: *C38, F63, O11, O15*

INTRODUCTION AND LITERATURE REVIEW

Is it possible to create a model that shows the measure of economic and social development for worldwide countries? Has this any connection with HDI? This represents the main idea for this paper, considering that HDI is a well- known index that consider variables like education, life expectancy and gross national income indexes. Considering indicators that are more oriented to economic, business, trade and labor development the authors try to reveal the connection between HDI that is already computed and known and new aggregated indicators, computed using analyses that reduce the dataset dimension.

The concern about HDI started years ago. In 2002, Biswas and Caliendo use a variables reduction analysis (and consider only one principal component) and the three indicators that compose HDI: gross domestic product per capita, life expectancy at birth and education for creating a similar indicator, named a metric for international human development. Their findings were similar to HDI and authors conclude that taking into

[1]* corresponding author, Lecturer, Phd, Romanian-American University, Bucharest, iordache_ana_mari_mihaela@yahoo.com
[2] associate professor, Phd, Bucharest University of Economic Studies

account the PCA is a more complex technique that brings more "straightforwardness" by generating optimally weights. On the other hand, Ranis, Stewart and Samman (2006) try to identify 11 categories of human development over 30 indicators. The categories are "mental well-being, empowerment, political freedom, social relations, community well-being, inequalities, work conditions, leisure conditions, political security, economic security and environmental conditions" (Ranis et. al., 2006). Authors used rank-order correlations among the variables from each category and identified the most relevant indicators for each category.

In 2004, Montenegro starts from the idea that, except income per capita, there is no rule to establish the most relevant variables that define the economic development. The author tries to define an economically developed country like a country with "high income per capita and a good income distribution" (Montenegro, 2004), where the terms high and good are understand differently by each person. After using the GDP per capita and Gini coefficients for a dataset of countries in order to develop an index, the author conclude with the recommendation to have a common methodology for Gini, recommendation for "United Nations, World Bank or the IMF" (Montenegro, 2004).

Later, in 2011, Abraham and Ahmed used data from 1975 to 2008, GDP as economic growth and HDI index as social development, in order to identify "the disequilibrium between the variables" in time using error correction model (ECM) as methodology. Estimating a regression model with GDP and HDI, the authors show a negative non-significant short-term relationship between these variables, but a very significant equilibrium coefficient for long-term relationship. So that the policies "aimed at accelerating growth would have a negative impact on human development in the short run but in the long run, equilibrium will be restored by HDI adjusting upwards or downward to correct the equilibrium error" (Abraham, Ahmed, 2011).

In 2014, Hajdouva, Z., Andrejovsky, P., Beslerova start from the idea that global experience does not confirm that economic development of countries comes with an increasing trend in quality of life. To do so, the authors chose 10 countries to study the relations "between the quality of life and environmental quality". By considering three clusters for all 10 countries, authors compare HDI with other indicators like corruption perception index (CPI), environmental performance index (EPI), GDP, and establish a model for future research that take into account indexes and variables like HDI, EPI, CPI and GDP.

The research divided into several sections: the introduction and literature review present the most relevant studies in this area of interest and the assumption that there is a connection between HDI and economic, business, trade and employment indicators, connection revealed by new indicators (components, factors). The methodologies section show briefly the statistical background for testing the assumption while the data selection and description is the part presenting the dataset used in the research. The last two parts is the result and interpretation where the main results are presented, and conclusions, with final details about this paper.

METHODOLOGIES

Principal components analysis (PCA) and factor analysis (FA) are two of the main dimension reduction analyses. Both have in common the idea of reducing the variables matrix dimension by keeping as much information as possible. But, the main difference between them is that PCA relies on an optimum problem that maximize the variance that each component take from all variable, while the factor analysis's idea is the assumption of the existing factorial model that, with a small number of factors, the patterns between correlated variables can be explained.

The model for principal components is (Dunteman, 1989):

- The first principal component is a linear combination of all X variables: $W_1 = \sum_{i=1}^{n} a_{1i} x_i$. The construction of this component takes into consideration that the variance of W (noted by λ) "is maximized by given the constraint that the sum of squared weights is one, and a_1 is an eigenvector associated to the first eigenvalue of the covariance matrix" (Dunteman, 1989).

- For the next principal component we should identify another eigenvector for the second eigenvalue, which maximize the variance of W. There is no correlation between it and the first principal component.

- This method continues until all n principal components are computed (as many as original variables), each of them having less variance and less information from X dataset than the previously component.

The idea of FA relies on[1]: "x = Λ f + e, for a p–element vector x, a p x k matrix Λ of loadings, a k–element vector f of scores and a p–element vector e of errors"[3]. If we consider the correlation matrix as Σ = Λ Λ' + Ψ, "the fit is done by optimizing the log likelihood assuming multivariate normality over the uniquenesses"[3]. Therefore, the scores might be written as f = Λ' Σ^{-1}x using Thomson's method, while "Bartlett's method minimizes the sum of squares of standardized errors over the choice of f"[3].

On the other side, the K-Means algorithm is an unsupervised learning algorithm that divides the dataset into a known number of classes taking into account the maximization of variance between classes and minimization the variance inside each individual class. This algorithm identifies one of the four classes of development, such as the HDI: low, medium, high and very high development.

Statistically, the steps for K-Means algorithm are[2]:

- Initially, we know the number of clusters. The k initial classes formed "randomly" with k observations within the data.

- each remained observation is then associated to one of the k clusters previously formed, using in general the method of the lowest distance between the initial centroid and each observation;

[1] https://stat.ethz.ch/R-manual/R-devel/library/stats/html/factanal.html
[2] https://en.wikipedia.org/wiki/K-means_clustering

- New centroids computed, after all observations grouped into k classes.
- "The algorithm repeats, until the convergence reached"[4].

In comparison with K-Means algorithm, the Ward's method (Ruxanda, 2009) is an ascending hierarchical classification method that takes into consideration the general criterion of classification: at each classification step, two classes that have the smallest sum of squares of deviations, comparing to other pairs of clusters. The idea behind this method is maximizing the homogeneity of clusters.

DATA SELECTION AND DESCRIPTION

The World Bank database is the main source of data. The indicators considered are for 2017 and reflect mostly the trade, employment, business and economic indicators that are relevant in analyzing the development degree of each country, and comparing the new results with HDI index. Therefore, the table with considered indicators (for 2017) and their codes is:

Table 1. Indicators used for models

Name [1]	Cod
"Age dependency ratio (% of working-age population)"[5]	I1
"Cost of business start-up procedures (% of GNI per capita)"[5]	I2
"Cost to export, border compliance (US$)"[5]	I3
"Cost to import, border compliance (US$)"[5]	I4
"Employers, total (% of total employment)"[5]	I5
"Employment in agriculture (% of total employment)"[5]	I6
"Employment in industry (% of total employment)"[5]	I7
"Employment in services (% of total employment)"[5]	I8
"GDP per capita growth (annual %)"[5]	I9
"Labor tax and contributions (% of commercial profits)"[5]	I10

Name[5]	Cod
"Merchandise exports (current US$)"[5]	I11
"Merchandise imports (current US$)"[5]	I12
"Merchandise trade (% of GDP) "[5]	I13
"Net migration"[5]	I14
"Population growth (annual %)"[5]	I15
"Profit tax (% of commercial profits) "[5]	I16
"Rural population growth (annual %)"[5]	I17
"Start-up procedures to register a business (number) "[5]	I18
"Tax payments (number) "[5]	I19
"Time required to start a business (days) "[5]	I20
"Urban population growth (annual %)"[5]	I21
"Wage and salaried workers, total (% of total employment)"[5]	I22

From 209 initial countries that had available data for 2017, only 144 left after outliers' removal. Each indicator has a name from I_1 to I_{22}, in the order mentioned above. All variables are standardized in order for using in further models.

[1] https://data.worldbank.org/

RESULTS AND INTERPRETATION

For creating aggregated indicators it is necessary some level of correlation between selected variables, so the first step is to identify the correlations between variables that can confirm the utility of both principal and factorial analyses.

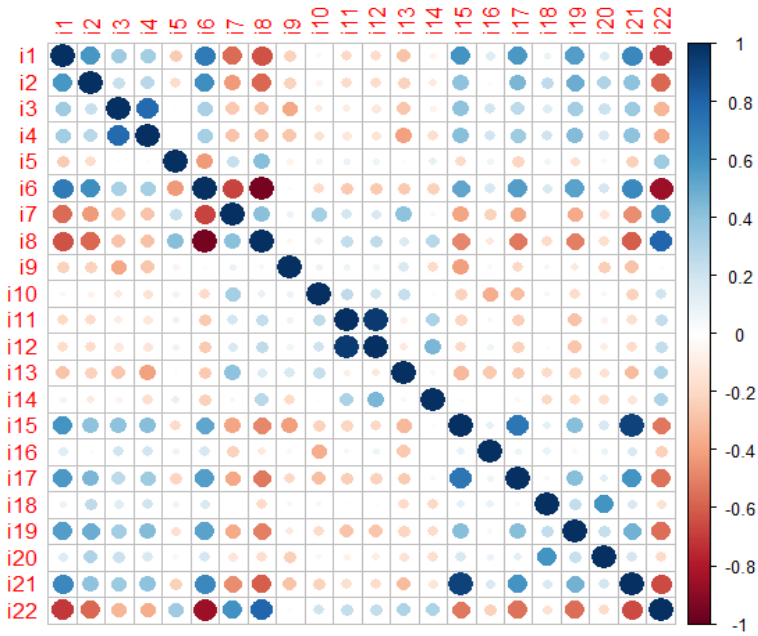


Figure 1. The correlation matrix between original variables

The figure from above is the correlation matrix between considered indicators. High correlations are between indicators like the employment in different areas (agriculture, industry and services) and the growth of urban population and I22. From these correlations it can be noticed that variables are splitting into two major components: the population component, that include variables like employment, population growth, labor tax and contribution, and a trade and economic component, including variables as the cost of export, import, gross domestic product per capita growth and number of tax payments. The correlation matrix from above represents that the dimension reduction analyses make sense, both analyses being methods to eliminate informational redundancy (no correlations between factors or principal components).

	lambda	ind_pr	cum_pr
Comp.1	7.48451	34.02052	34.02052
Comp.2	2.24158	10.18902	44.20954
Comp.3	1.90097	8.64080	52.85033
Comp.4	1.63630	7.43772	60.28805
Comp.5	1.43491	6.52230	66.81035
Comp.6	1.14815	5.21885	72.02919
Comp.7	0.86850	3.94772	75.97691

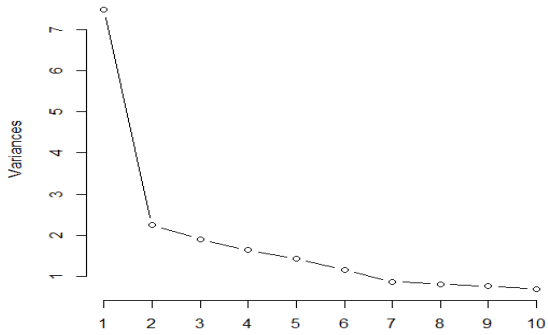


Figure 2. PCA statistics

The figure from above represents the main principal components analysis statistics. According to Kaiser criteria of choosing the number of principal components (PCs), 6 of 22 components may be considering for model, because of a variance higher than one, fact that is confirmed by the above scree plot: starting with component 7, the slope becomes almost insignificant and the amount of information brought by each new component decreases very much. Therefore, from 100% of information, 72% contained by first six components are enough to provide relevant conclusions.

```
> res <- paf(as.matrix(date_std))
> res$KMO
[1] 0.71558
> res$Bartlett
[1] 4946.9
> qchisq(0.05,(22*21/2),lower.tail=F)
[1] 267.45
```

	Factor1	Factor2	Factor3	Factor4	Factor5
SS loadings	4.27	2.32	2.27	2.21	1.48
Proportion Var	0.19	0.11	0.10	0.10	0.07
Cumulative Var	0.19	0.30	0.40	0.50	0.57

Figure 3. FA statistics

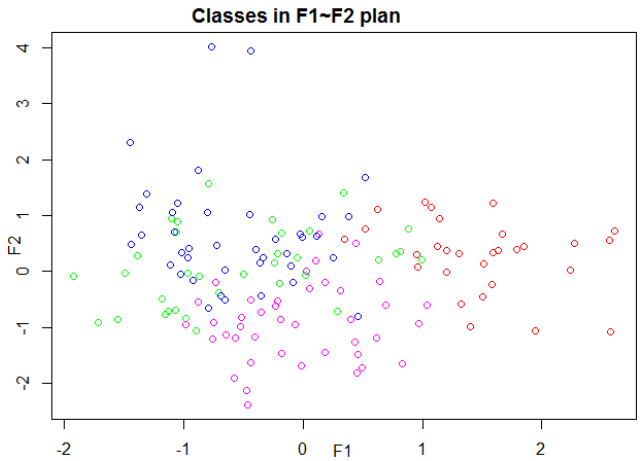
If PCA is a way to reduce the number of correlated variables, the FA idea is that there is a factorial model that fits to a number of variables. In this respect, a KMO (Kaiser-Meyer-Olkin) value calculated for standardized variables has a value closer to 1 (0.71) and show the utility of using FA. Moreover, the Bartlett's Sphericity test show the rejection of null

hypothesis. This means that the correlation matrix is equal to unity, so the variables are orthogonal, without correlations. Both KMO and Bartlett's test show that the assumption that a factorial model exists between variables is true, so that the chosen number of five factors (57% of total variability) can explain the patterns between variables.

```
"W1" "W2" "W3" "W4" "W5" "W6"
"1" -3.7239 -1.2945 -1.0491 0.2113 -0.1861 -0.941
"2" -3.6055 0.7017 1.3462 -0.0179 -0.3905 0.2991
"3" 0.3254 -0.1405 -0.9163 0.0321 0.2872 0.1625
"4" 2.9282 0.15 0.7717 -0.0981 -0.106 -0.1338
   "F1" "F2" "F3" "F4" "F5"
   "1" 1.4863 0.2862 -0.1133 -0.1731 -0.6028
   "2" -0.5794 0.6685 0.5701 -0.4753 0.0384
   "3" -0.5268 0.0804 -0.2795 1.144 -0.3684
   "4" -0.0491 -0.9214 -0.2443 -0.3626 0.6945
```

Figure 4. Classes centroids using K-Means

The centroids for each of four classes resulting from K-Means algorithm have interpreted in terms of factors and principal components. Each factor or component "takes" information from all 22 variables selected (more or less, depending on coefficients or eigenvectors), but, taking into account the correlations between observed variables and computed new aggregates (factors or components), it is easy to determine the name of each class, based on the average values from above. In this respect, the classes for principal components analysis (W) are (as level of development) 1= high, 2=low, 3=medium, 4=very high, while the classes for extracted factors are (as level of development) 1=low, 2=medium, 3=high, 4=very high.



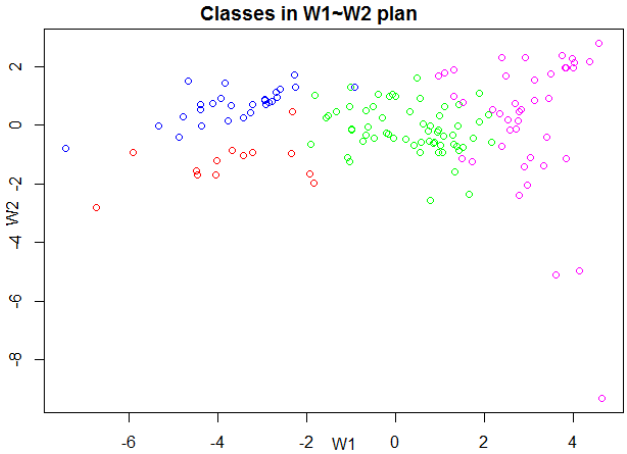


Figure 5. Classes (K-Means classification) representation in factorial (left) and principal (right) plan

The figure from above show the graphically representation of all four classes in two dimensions: the first two factors (left side of the figure) and the first two PCs (right part of the figure). Even if the amount of information explained by first two factors or taken by first two PCs is not very high, comparing to five factors and six components, the clarity with which the four classes are distinguished is remarkable.

```
> confMatrix_w
                                Predicted
Original                          1  2  3  4
  HIGH_HUMAN_DEVELOPMENT         17  0 10  8
  LOW_HUMAN_DEVELOPMENT           1 28  1  0
  MEDIUM_HUMAN_DEVELOPMENT        6  5 19  0
  VERY HIGH HUMAN DEVELOPMENT    15  0  0 30
> confMatrix_f
                                Predicted
Original                          1  2  3  4
  HIGH_HUMAN_DEVELOPMENT          0  8 13 14
  LOW_HUMAN_DEVELOPMENT          21  2  7  0
  MEDIUM_HUMAN_DEVELOPMENT        8  6  8  8
  VERY_HIGH_HUMAN_DEVELOPMENT     0 23  4 18
```

Figure 6. Confusion matrix for factors and principal components using K-Means

Confusion matrix are the most important outputs, because it represents the connection between the class showed by HDI, also named as original class and the new class (predicted) computed using K-Means algorithm and new datasets: principal components (W) and factors (F). From this point of view, knowing the new class signification from above, it is possible to estimate the "accuracy" degree, similar to supervised learning techniques, but with a different signification here. If we consider six principal components, then, there is an approximate 67% connection between new variables

classification output and HDI classification. On the other side, using factors, the percentage is only about 42%. The difference among the results from above comes from the amount of total information that each dataset takes from original variables, as well as the method applied to reduce data dimensionality: the difference between an optimum problem and maximum likelihood estimation.

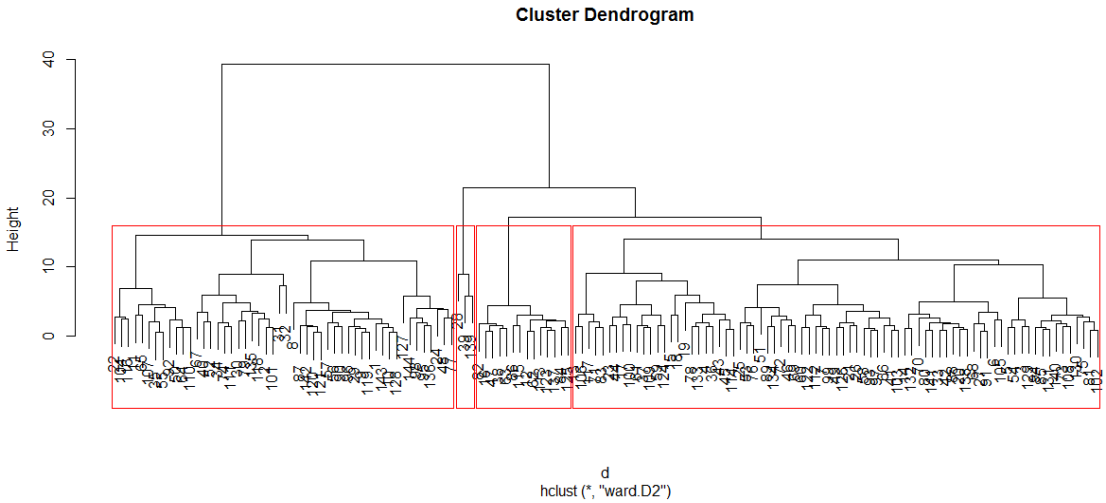


Figure 1. Ward's dendrogram using principal components

The figure from above show the dendrogram obtained by applying the Ward's hierarchical method on principal components. The graph using red squares represents the fourth classes.

```
> confMatrix_f2
                                Predicted
Original                          1  2  3  4
  HIGH_HUMAN_DEVELOPMENT          6 25  3  1
  LOW_HUMAN_DEVELOPMENT          25  4  1  0
  MEDIUM_HUMAN_DEVELOPMENT       13 17  0  0
  VERY_HIGH_HUMAN_DEVELOPMENT     0 42  2  1
> confMatrix_w2
                                Predicted
Original                          1  2  3  4
  HIGH_HUMAN_DEVELOPMENT          2 30  2  1
  LOW_HUMAN_DEVELOPMENT          30  0  0  0
  MEDIUM_HUMAN_DEVELOPMENT       18 12  0  0
  VERY_HIGH_HUMAN_DEVELOPMENT     0 32 11  2
```

Figure 8. Confusion matrix for factors and principal components using Ward

It is interesting to compare the classification methods like K-Means and Ward's method from confusion matrix point of view. Theoretically, K-Means provides better results of

classification, because is an algorithm that runs until a stop condition is fulfilled, but, Ward's classification method is known as providing similar classification results as an algorithm. In this respect, we presented above the confusion matrixes, for both principal components (w2) and factors (f2). According to classes' centroids and factors/components meaning, the new names for classes are 1=low, 2=high, 3=very high, 4=medium development level, for principal components. The percent of "correct classification" is about 51% (lower than a K-Means classification), and 1=low, 2=very high, 3=medium, 4=high development for factors and the percent is about 49%, higher than K-Means classification.

CONCLUSIONS

Finally, the conclusion of the article is to demonstrate the connection between the HDI index and the most relevant indicators from trade, employment, business and economic, taking into consideration the majority of worldwide countries. The dimension reduction methods (PCA and FA) were used to both synthesize the information from 22 variables and to create new indicators. Further analyses used these new indicators. Both K-means algorithm and Ward's method were used to classify the 144 countries into 4 classes and then to compare the new obtained classes with HDI, in order to see the connection between HDI index and new proposed models. In this respect, the confusion matrix estimate this connection in terms of "correct" classification percentage. For further analyses, we propose to analyze the degree of development for each group of countries by including more than 22 indicators, like social, education, health, poverty or financial indicators.

REFERENCES

[1] Abraham, T.W., Ahmed, U.A. (2011), Economic Growth and Human Development Index in Nigeria: An Error Correction Model Approach, International Journal of Administration and Development Studies, University of Maiduguri, Nigeria, Vol. 2, No. 1, pp. 239-254, ISSN: 2141-5226"

[2] Biswas, B., Caliendo, F. (2002), A Multivariate Analysis of the Human Development Index, Economic Research Institute Study Papers, Paper 244."

[3] Dunteman, G.H. (1989), Principal Components Analysis, Ed. SAGE, 1989, ISBN 0803931042, 9780803931046."

[4] Hajdouva, Z., Andrejovsky, P., Beslerova, S. (2014), Development of quality of life economic indicators with regard to the environment, Procedia - Social and Behavioral Sciences, Vol. 110 (2014), pp. 747 – 754

[5] Montenegro, A. (2004), An Economic Development Index, Development and Comp Systems, University Library of Munich, Germany, https://EconPapers.repec.org/RePEc:wpa:wuwpdc:0404010."

[6] Ranis, G., Stewart, F., Samman, E. (2006), Human Development: Beyond the Human Development Index, Journal of Human Development, Vol. 7, No. 3, pp. 323-358."

[7] Ruxanda, Ghe. (2009), Analiza multidimensionala a datelor, Academia de Studii Economice, Scoala Doctorala, Bucuresti, 2009

[8] http://databank.worldbank.org/data/home.aspx

[9] https://data.worldbank.org/

[10] https://en.wikipedia.org/wiki/List_of_countries_by_Human_Development_Index"

[11] http://hdr.undp.org/en/composite/HDI"

[12] https://stat.ethz.ch/R-manual/R-devel/library/stats/html/factanal.html"

[13] https://en.wikipedia.org/wiki/K-means_clustering"