Pak. J. Statist.
2021 Vol. 37(2), 101-119

PRINCIPAL COMPONENT ANALYSIS AND APPLICATION TO PUBLIC EXPENDITURE EFFICIENCY INDICATORS

Baraka Achraf Chakir[1§], Driss Mentagui[1], Ahssaine Bourakadi[1] and **Yamoul Nada[2]**

[1] Department of Mathematics, Faculty of Sciences, Ibn Tofail University Kenitra City, Morocco.

[2] Department of Physics, Faculty of Sciences, Ibn Tofail University Kenitra City, Morocco.

[§] Corresponding author Email: baraka.achraf.chakir@gmail.com

ABSTRACT

The Multicriteria classification and data processing of quantitative and qualitative indicators cannot be carried out directly by conventional statistical methods. In this work, the principal component analysis of the indicators of the efficiency social spending following the transformation of the variables (centered, normalized), was carried out through projections of point clouds in the space of the plans factorials. Corresponding to multivariate analyzes of variance. This analysis made it possible to classify the effectiveness of social spending in 14 emerging countries.

Subsequently, a comparison with the ascending hierarchical classification using the Ward method confirms the rankings obtained, ranking Morocco among the less efficient in terms of social spending countries at the same level as most of the North African countries. The results come from R and SPSS software.

KEYWORDS

Principal component Analysis, Hierarchical Ascending Classification, SPSS and R software.

INTRODUCTION

Principal Component Analysis (PCA) is an extremely powerful information synthesis tool, very useful when analyzing a large amount of quantitative data to be processed and interpreted. These methods first appeared in the early 1930s by Hotteling, and then developed in France in the early 1970s, by the mathematician Jean-Paul Benzecri who analyzed geometric aspects and graphic representations on the different factor axes.

The PCA depends on a geometric model. It is not based on probabilistic methods as is the case for a set of data analysis tools. The PCA proposes, from a data table made up of the values of p quantitative variables and n units (also called individuals), geometric representations of all the units and variables.

There are many statistical techniques to divide a population into different classes or subgroups. One of them is the ascending hierarchical classification (CAH). This

technique aims to group together individuals who have the same characteristics within the same class (intra-class homogeneity).By increasing the possibility that the classes are the most dissimilar (inter-class heterogeneity).

The principle of the CAH is to gather individuals according to a criterion of resemblance defined beforehand which will be expressed in the form of a matrix of distances, expressing the distance existing between each individual. Two identical observations will have a zero distance. The more dissimilar the two observations, the greater the distance.

The realization of the importance of social development has led governments to reconsider the efficiency of public spending in the social fields. These expenditures are intended to meet the basic needs of the populations in addition to the growth of the national income. Given the many effects of social spending, their impacts are both difficult to identify and have multiple horizons.

In this sense, we proceed to the classification of the efficiency of social spending [1], of 14 emerging countries, according to several methods: first by principal component analysis (PCA) This last synthesizes the data by constructing a small number of new variables, the main components. The essential elements of the data table can then be entered quickly, using graphic representations established from these main components. Thereafter, we will use the ascending hierarchical classification which is based on the wad method In order to compare the classifications obtained from these two methods. All calculations and figures come from R and SPSS software.

1. THEORETICAL PART OF THE PRINCIPAL COMPONENT ANALYSIS

1.1 Presentation of the Data Table

The data is represented in the form of a crosstab whose measurements performed on n units $\{u_1, u_2, \dots u_i, \dots u_n\}$. And p quantitative variables are represented by the measurements $\{v_1, v_2, \dots, v_j, \dots v_p\}$. The data table on which the analysis is based is denoted by X and is written in the following form:

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}$$

Each unit can be represented by the vectors of the measurements on the p variables.

$$t_{U_i} = (x_{i1}, x_{i2}, \dots x_{ip})$$

All the variables of the matrix X can be represented by a vector R^n whose components are the values of the variable for the n units.

$$V_j = \begin{bmatrix} x_{1j} \\ x_{2j} \\ . \\ x_{ij} \\ . \\ x_{nj} \end{bmatrix}$$

To have an image of the set of units, instead in a refined space as an origin as a particular vector of R^p, for example, the vector of which all the coordinates have a zero value. Thereafter, each unit will be represented by a point in this space. The set of points which represent the totality of the units is known as the "cloud of individuals".

Using the same procedure in R^n, each variable can also be represented by a point in space. The set of points representing the variables is considered as a cloud, hence the name "variable cloud".

We note that these spaces being of dimension greater than 3 in general, which gives the possibility of visualizing the representations of the clouds of individuals and variables. The general idea of factorial methods is to fix a system of axes and planes such as the projections of these point clouds on these axes and these planes make it possible to construct classes which have the same characteristics. That is, to have the least distorted images.

1.2 The Choice of the Distance between Two Units

In order to choose a geometric representation, choosing a distance between two points in space is essential [2]. The classical Euclidean distance is the distance used by PCA in the corresponding affine space. The distance between two units u_i and $u_{i"}$ is equal to:

$$d^2(u_i, u_{i"}) = \sum_{i=1}^{p} (x_{ij-} x_{i'j}).$$

Using distance, the axes defined by the variables form an orthogonal basis. At this distance we associate a scalar product between two vectors:

$$< \overrightarrow{OU_\iota}, \overrightarrow{OU_{\iota'}} >= \sum_{J=1}^{p} x_{ij}\, x_{i'j} = {}^tU_i U_{i'}$$

So the norm of a vector is defined by:

$$\overrightarrow{||\, OU_\iota}||^2 = \sum_{J=1}^{p} {x_{ij}}^2 = {}^tU_i U_i.$$

The angle α between two vectors is defined by:

$$\cos(\alpha) = \frac{< \overrightarrow{OU_\iota}, \overrightarrow{OU_{\iota'}} >}{||\, \overrightarrow{OU_\iota}||^2 \overrightarrow{||OU_{\iota"}||^2}} = \frac{\sum_{J=1}^{p} x_{ij}\, x_{i'j}}{\sqrt{\sum_{J=1}^{p} {x_{ij}}^2 \sum_{J=1}^{p} {x_{i'j}}^2}} = \frac{{}^tU_i U_{i'}}{\sqrt{{}^tU_i U_i\ {}^tU_{i'} U_{i'}}}.$$

1.3 Choice of Origin

The point "o" corresponding to the vector whose coordinates are all zero, cannot be considered as a satisfactory origin. The coordinates of the points in the cloud of individuals control the position of the cloud for example if the coordinates are large, the cloud will be positioned far from the origin [3]. It seems wiser to choose an origin that

will be linked directly to the point cloud: the center of gravity of the cloud. In order to define this center of gravity, the choice of the unit weighting system must satisfy:

$\forall\, i = 1, \ldots, n, P_i =$ weight of unit U_i, such as $\sum_{i=1}^{n} P_i = 1$.

The center of gravity is defined as being the point which verifies:

$$\sum_{i=1}^{n} P_i \ \overrightarrow{GU_i} = \vec{0}$$

For PCA, the weight $\frac{1}{n}$ is assigned to all individuals .The center of gravity G of the cloud of individuals is then the point whose coordinates are the mean values of the variables:

$$G = \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n} x_{i1} \\ \vdots \\ \frac{1}{n}\sum_{i=1}^{n} x_{ij} \\ \vdots \\ \frac{1}{n}\sum_{i=1}^{n} x_{ip} \end{bmatrix} = \begin{bmatrix} x_{.1} \\ \vdots \\ x_{.j} \\ \vdots \\ x_{.p} \end{bmatrix}$$

Taking G as the origin, in accordance with the following figure, then amounts to working on the table of centered data:

$$X_c = \begin{bmatrix} x_{11} - x_{.1} & \cdots & x_{1p} - x_{.p} \\ \vdots & \ddots & \vdots \\ x_{n1} - x_{.1} & \cdots & x_{np} - x_{.p} \end{bmatrix}$$

The vector of the centered coordinates of the unit U_i is written in the form:

$$U_{ci} = \begin{bmatrix} x_{i1} - x_{.1} \\ \vdots \\ x_{ip} - x_{.p} \end{bmatrix}$$

That of the centered coordinates of the variable v_i is:

$$V_{cj} = \begin{bmatrix} x_{1j} - x_{.j} \\ \vdots \\ x_{nj} - x_{.j} \end{bmatrix}$$

1.4 Moments of Inertia

Total inertia of the cloud of individuals: we note I_G the moment of inertia of the cloud of individuals with respect to the center of gravity G:

$$I_G = \frac{1}{n}\sum_{i=1}^{n} d^2\left(G, U_i\right) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_{ij} - x_{.j})^2 = \frac{1}{n}\sum_{i=1}^{n} {}^tU_{ci}\, U_{ci}$$

The dispersion of the cloud of individuals can be measured by the total moment of inertia, which reflects the importance of this measurement of the cloud of individuals relative to its center of gravity. The total dispersion of the cloud means that the moment of inertia is very large, while if it is small, the cloud concentrates on its center of gravity, which marks less dispersion.

By inverting the order of the sums, I_G is written in the following form:

$$I_G = \sum_{j=1}^{p}\left(\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - x_{.j})^2\right) = \sum_{j=1}^{p} Var(v_i).$$

where $Var(v_j)$ is the empirical variance of the variable V_j. In this form, we see that the total inertia is equal the trace of the covariance matrix Σ (cf. appendix A) of the p variables V_j:

$$I_G = trace\ (\Sigma)$$

Inertia of the cloud of individuals with respect to an axis passing through G: The inertia of the cloud of individuals with respect to an axis Δ passing through G is equal, by definition

$$I_\Delta = \frac{1}{n}\sum_{i=1}^{n} d(h_{\Delta i}, u_i)^2.$$

where $h_{\Delta i}$ is the orthogonal projection of u_i on the axis Δ. This inertia measures the proximity to the Δ axis of the cloud of individuals.

The inertia of the cloud of individuals with respect to a vector subspace V passing through G: This inertia is, by definition, equal to:

$$\mathrm{I_V} = \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{d}(\mathrm{h_{Vi}}, \mathrm{u_i})^2.$$

where $\mathrm{h_{Vi}}$ is the orthogonal projection of $\mathrm{u_i}$ on the subspace V, one seeks the axis Δ_1 which passes directly through the point G and minimizes the inertia.

Search for axis $\boldsymbol{\Delta_1}$ passing through point G of minimum inertia: By looking for a passing axis Δ_1 passing through G of inertia I_{Δ_1} minimum because it is the axis closest to all the points of the cloud of individuals, therefore the projection of this cloud on this axis is the one which gives the least distorted image of the cloud. If using the relationship between the properties given in the previous paragraph, finding Δ_1 for the minimum that: I_{Δ_1} is minimum is equivalent to finding Δ_1 such that $I_{\Delta"_1}$ is maximized.

$I_{\Delta 1}$ is minimum $\Leftarrow\Rightarrow$ $I_{\Delta^* 1}$ is maximum.

the axis Δ1 defines by its unit director vector $\overrightarrow{Ga_1}$, we must therefore find $\overrightarrow{Ga_1}$ such that is I_{Δ^*1} maximum under the constraint that $\overrightarrow{Ga_1}$. = 1.

1.5 Maximum Search

The problem to solve: finding a_1 such that ${}^t a_1 \Sigma\, a_1$ is maximum with the constraint ${}^t a_1\, a_1 = 1$, is the problem of finding an optimum of a function of several variables linked by a constraint (the unknowns are the components of a_1). The Lagrange multiplier method can then be used. In the case of searching for a_1, you have to calculate the partial derivatives of:

$$g(a_1) = g(a_{11}, a_{12}, \ldots, a_{1p}) = {}^t a_1 \Sigma a_1 - \lambda_1({}^t a_1\, a_1 - 1).$$

Using the matrix derivative, on:

$$\frac{dg(a_1)}{da_1} = 2\,\Sigma\, a_1 - 2\lambda_1 a_1 = 0.$$

The system to solve is:

$$\begin{cases} \Sigma\, a_1 - \lambda_1 a_1 = 0 \ (1). \\ {}^t a_1\, a_1 - 1 = 0 \ (2) \end{cases}$$

From the matrix equation (1) of this system, we deduce that $a_1$is the eigenvector of the matrix Σ associated with the eigenvalue λ_1. By multiplying on the left by ${}^t a_1$ the two members of equation (1) we obtain:

$${}^t a_1\, \Sigma\, a_1 - \lambda_1 {}^t a_1 a_1 = 0$$

And using equation (2) we find that:

$${}^t a_1 \Sigma\, a_1 = \lambda_1.$$

It is recognized that the first member of the previous equation is equal to the inertia which must be maximum. This means that the eigenvalue is the largest eigenvalue of the covariance matrix and that this eigenvalue is equal to the inertia carried by the axis Δ_1.

The axis Δ_1, for which the cloud of individuals has minimum inertia as a unitary Director vector the first eigenvector associated with the largest eigenvalue of the covariance matrix Σ.

2. PRINCIPAL COMPONENT ANALYSIS OF THE SOCIAL SPENDING EFFICIENCY INDICATORS USED:

2.1 Formulation of the Problem

In this part, we will study the situation of 14 emerging countries which had a similar economic situation in the early 70s, belonging to different continents. We take a set of 12 social expenditure efficiency indicators as variables, we will apply the central limit theorem [6], to these indicators (centered, normalized).

The 12 social indicators relate to several dimensions, namely: education, health, security, innovation, the labor market and poverty. All the results of this part come from the software *R*[4], data are from World Bank site [12].

Table 1
Social Spending Efficiency Indicators Used

Field	Indicators
Education	Youth employment rate (YER)
	Youth unemployment rate (YUR)
Health	Life expectancy (LE)
	Number of hospital beds by 1000 inhabitants (NBI).
	Number of doctors by 1000 inhabitants (NDI)
Security	Number of murders by 10.000 inhabitants (NMUI)
	Number of prisoners by 100.000 inhabitants (NPRI)
	The share of military spending in GDP (MS)
Innovation	The share of R&D (in% GDP) (R&D)
	Number of triadic patents (% population) (NTP)
Poverty	Percentage of the population living on at least $ 2 a day (PP)
Labor Market	Employment rate (ER)

Source: Author.

2.2 Correlation Table

The Table presents the correlation matrix of the 12 study variables. This step is necessary to detect the strong and weak correlations before applying the PCA (centered, normalized) of all the variables.

Table 2
La The Correlation Matrix of the 12 Variables

Correlation matrix												
	YER	YUR	LE	NBI	NDI	NMI	MS	NPRI	R&D	NTP	PP	ER
YER	1,000	-0,761	-,170	-,055	-,146	,787	-,411	,125	,156	,610	,192	,880
YUR		1,000	-,183	,228	,009	-,539	,541	-,041	,289	-,136	,277	-,840
LE			1,000	,289	,141	-,001	-,415	,562	,000	-,192	-,300	,122
NBI				1,000	,148	,278	-,292	,087	,578	,533	-,184	-,075
NDI					1,000	,177	-,117	-,266	-,358	,065	-,250	-,158
NMUI						1,000	-,491	,169	,187	,800	-,001	,740
MS							1,000	-,257	-,003	-,194	,329	-,442
NPRI								1,000	,547	,163	,201	,240
R&D									1,000	,574	,285	-,021
NTP										1,000	,110	,476
PP											1,000	,065
ER												1,000

Source: Author

The correlation matrix suggests that there are negative linear correlations, which means that these variables vary in opposite directions, for example that between (the youth unemployment rate and the youth employment rate) and (l 'life expectancy and youth unemployment rate) coefficient respectively -0.761, -0.183.

We find a strong negative correlation between the youth employment rate and the youth unemployment rate, thus between the youth unemployment rate and the employment rate. While the rest of the variables are positively correlated which shows that these variables vary in the same direction, some being strong (0.88 and -0.840), other means (0.578; 0.562 and -0.442), others have a very weak correlation (-0.003, -0.021, 0.110).

Bartlett's Sphericity test: To better ensure that the PCA is valid, we must use the sphericity test, interpretation of the test:

Indice KMO et test de Bartlett		
Precision measurement of Kaiser-Meyer-Olkin sampling.		0,566
Bartlett's sphericity test	Approximate chi-square	117,348
	Ddl	66
	Meaning of Bartlett	0,000

Source: Author

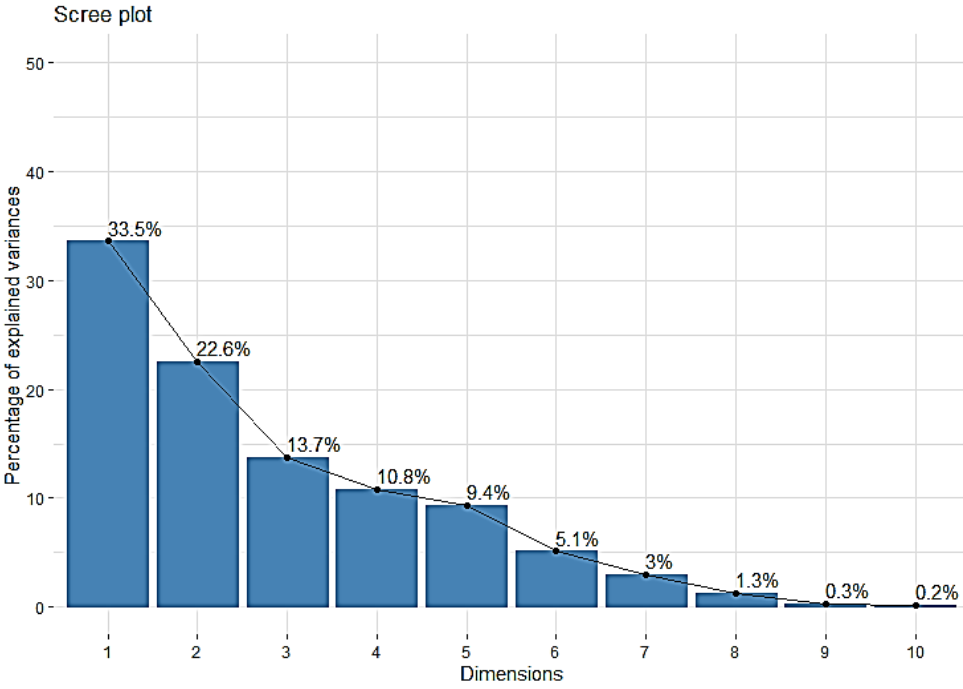
Figure 1: The KMO Index and Bartlett's Sphericity Test

H0: there is no significant correlation between the variables.
H1: At least one of the correlations between the variables is significant.

Given the error of the meaning = 0.00 < 0.05 we reject the null hypothesis H0, so the test is significant, therefore we can perform a principal component analysis without any problem.

2.3 Variance table explained:

The variance of a main component [8] ,is equal to the inertia carried by the main axis associated with it.



Source: Author

Figure 2: Eigenvalue Graph

We can see from the graph that the first factor has an eigenvalue of 4.109; that is to say that it carries a quantity of 4,109 information, that is to say almost 33.5% of the total information present in the table, that is to say that if the we represent the data on a single axis, We will always have 34.246% of the total variability which will be preserved. The second is less informative with an amount of information about 2,369 or almost 22.6% of the total information and for the third factor which represents 2,023 of the information or almost 13.7% of the total information. The fourth axis contributes by 10.7% to explain the total variation, with a quantity of information 1.61.

As the factors (5 to 12) do not explain enough variance, they are not retained, we see that the first four eigenvalues are greater than 1. This brings us back to take into account the four corresponding factor axes. In addition, the first factorial axis (F1) alone explains 34.246% of the total inertia, the second (F2) and the third and the fourth (F3) and (F4) explain 19.745%; 16.858% and 13.417%. The four axes which have been mentioned explain 81% of the total variation.

In order to justify the choice to take into account the first factorial plane (F1, F2, F3, F4), we have drawn the graph of the eigenvalues. The graph confirms the decision previously made to keep the four axes F1, F2, F3, F4. Indeed, there is an inflection of the curve at the level of the fourth eigenvalue.

2.4 The Components Matrix

Each axis is associated with a variable called the main component. Component 1 is the vector containing the coordinates of the projections of individuals on the axis 1. Component 2 is the vector containing the coordinates of the projections of individuals on the axis 2. To obtains these coordinates, we write each principal component as a linear combination of the initial variables.

Table 3
Component Matrix

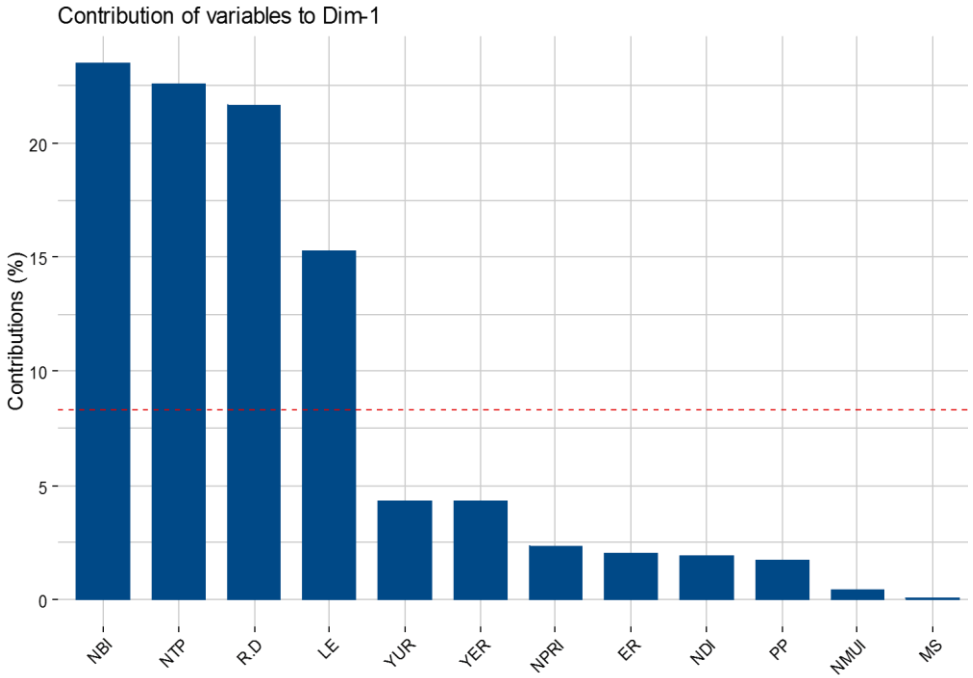
Variables	Components			
	1	2	3	4
Number of murders by 10.000 inhabitants	0,902	0,023	-0,04	0,29
Youth employment rate	0,89	-0,12	-0,38	-0,006
Employment rate	0,88	-0,25	-0,19	-0,18
Youth unemployment rate	-0,72	0,60	0,01	0,17
Number of triadic patents	0,71	0,47	-0,12	0,44
The share of military spending in GDP	-0,62	0,20	-0,50	0,07
The share of R&D (in% GDP)	0,265	0,91	,010	-0,07
Number of hospital beds by 1000 inhabitants	0,24	0,59	0,53	0,41
Life expectancy	0,15	-0,02	0,80	-0,44
Percentage of the population living on at least $ 2 a day	-0,01	0,41	-0,62	-0,23
Number of prisoners by 100.000 inhabitants	0,33	0,45	0,27	-0,67
Number of doctors by 1000 inhabitants	-0,03	-0,28	0,39	0,60

Source: Author.

From this table, it appears that the variables "Number of murders by 10,000 inhabitants" and "Youth employment rate and Employment rate" and "Youth unemployment rate" and "Number of triadic patents" and "The share of military spending in GDP" contribute more than the others to the inertia explained by the axis F1. We note that these six variables are the best represented on this axis. The same remark is noted for the variable "The share of R&D (in% GDP)", "Number of hospital beds by 1000 inhabitants" concerning the axis F2 and for the variables "Life expectancy" and "Percentage of the population living on at least $ 2 a day" contribute to the inertia explained by the axis F3 and finally the two variables "Number of hospital beds by 1000 inhabitants" and "Number of doctors by 1000 inhabitants" for axis F4.

2.6 Analysis of the First Factorial Axis (F1)

To give an interpretation to a factorial axis, one seeks the variables having a good quality of representation and whose contributions are the strongest while distinguishing those of the positive coordinates and those of the negative coordinates.



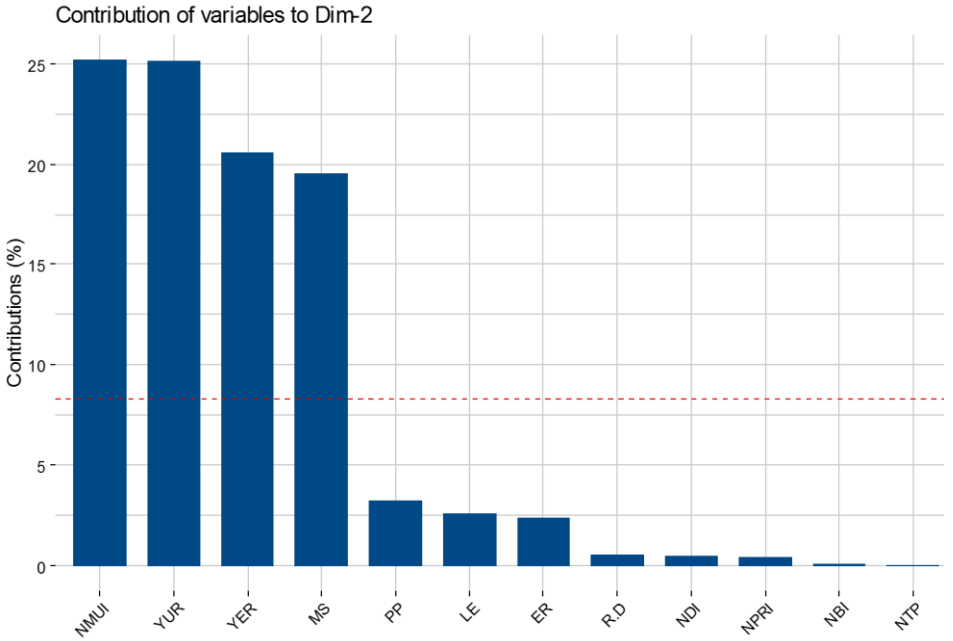
Source: Author

Figure 3: Variables Graph by Axis F1

From the graph, it appears that the variables "Number of hospital beds by 1000 inhabitants", "Number of triadic patents", "The share of R&D (in% GDP)", contribute more than the others to the inertia explained by the F1 axis. Variables forming the elements of these sets contribute more than 70% to the inertia explained by this axis.

2.5 Analysis of the Second Factorial Axis (F2)

From the graphic, we can determine the variables that contribute the most to the inertia of axis F_2:



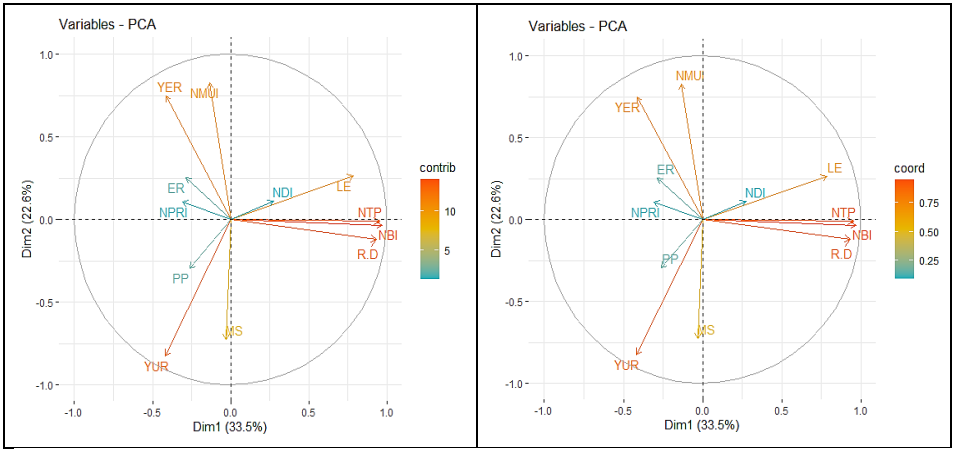
Source: Author

Figure 4: Variables Graph by Axis F2

The contribution of all these variables to the explanation of axis F2 amounts to more than 62%, we see for the variables "Number of murders" by 10,000 inhabitants "Youth unemployment rate" and "Youth employment rate" concerning the F2 axis, contribute more than in inertia explained by this axis.

2.6 Analysis in the First Factorial Plane (F1, F2)

In terms of the indicators: In this context, we recall that the first factorial plan explains more than 57% of the information contained. The graphic representation of the variables on this plane allows, at a glance, to visualize the indicators which contribute positively and negatively at the level of the two axes as shown in the following factor map.

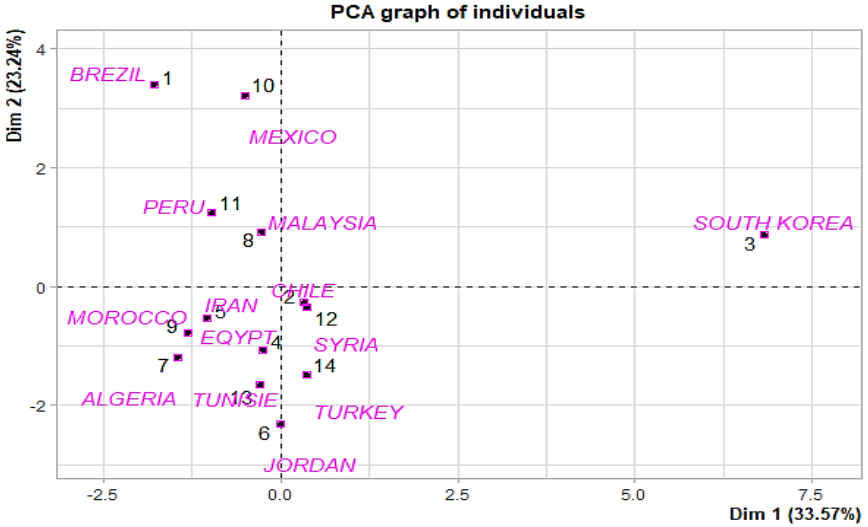


Source: Author

Figure 5: Graphical Representation of the Indicators in the Factorial Plane (F1, F2)

The presentation of the indicators by coordinates and by contribution on the map shows that the indicators which act positively: "The share of military expenditure in the GDP", "Number of doctors by 1000 inhabitants", "Life expectancy" and " Youth employment rate "point in the opposite direction to axis F2. While the Percent of population living on at least $2 a day variables, "Youth unemployment rate", "Number of murders by 10000 inhabitants", act negatively in the other direction for this axis. This indicates that there is a negative correlation between these two groups of variables and a positive correlation between the variables within each group.

The graphical representation of the cloud of individuals on the factorial plane (F1, F2) [9] illustrates the position of the countries in relation to each of the axes. It also makes it possible to group individuals with similar characteristics. The projection of the country points on the map (F1, F2) is illustrated by the following.

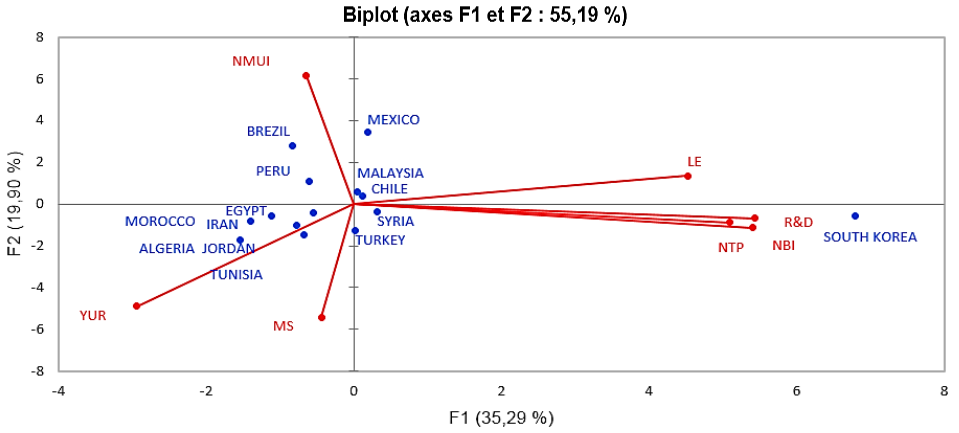


Source: Author

Figure 6: Graphic Representation of Countries in the Factor Plane (F1, F2)

The proximity of two country points in a factorial plane means that these two countries have a similar behavior towards all the variables. Indeed, in the first factorial plane (F1, F2), we can build three homogeneous groupings of regions according to their degree of performance. From this figure, we note that there are three sets:

{Mexico, Peru, Malaysia, Brazil}
{Syria, Morocco, Algeria, Jordan, Egypt, Iran, Turkey, Tunisia, Chile}
{South Korea}.



Source: Author

Figure 7: Graphical Representation of the Indicators, and of the Countries in the Factorial Plane (F1, F2)

The first factorial axis is represented by South Korea, which has strong correlations with the indicators "Life expectancy", "Number of beds by 1000 inhabitants", "Number of triadic patents", "The share of R&D (in% GDP)". This reflects the place occupied by research and development and efforts to improve the financing of its activities, both on the part of the State and of the private sector in this country.

Latin American countries do not see difficulties in terms of integration into the labor market, but there are problems concerning the security of individuals, while for most the Arab countries of North Africa have educational indicators, low sanitary compared to other countries.

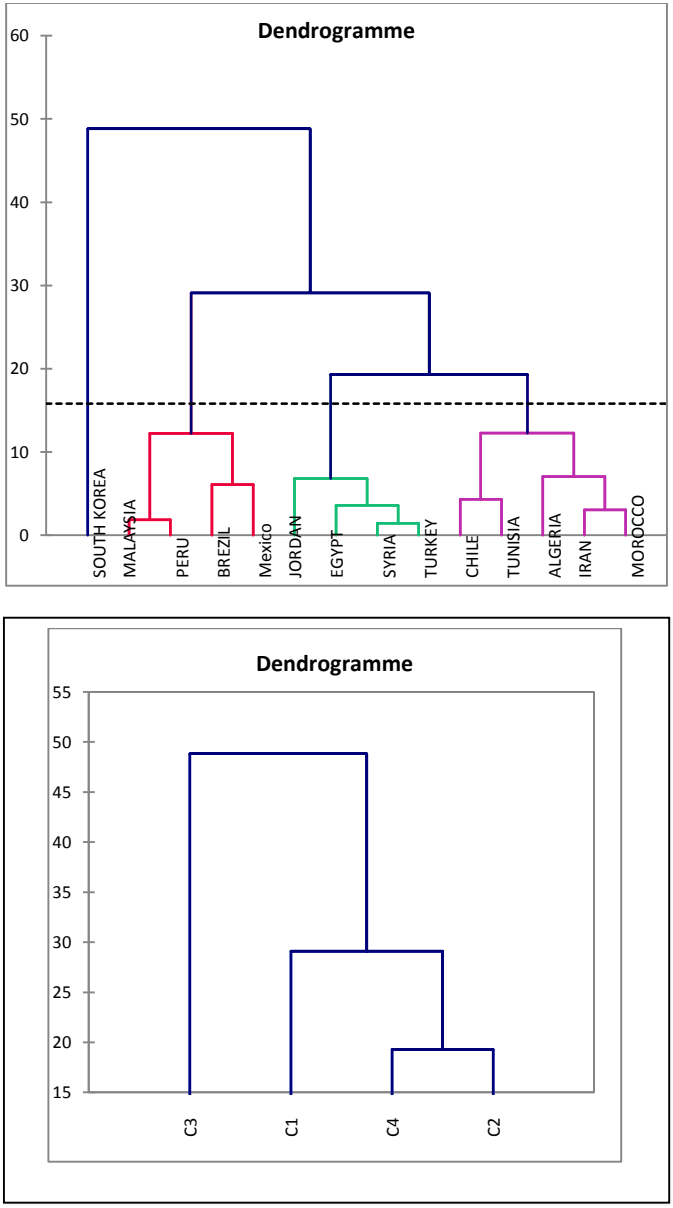
3. COMPARISON BETWEEN PRINCIPAL COMPONENT ANALYSIS AND HIERARCHICAL ASCENDING CLASSIFICATION

The factorial axes (F1, F2, F3, F4) explain 81% of the total variance, consequently. The groupings proposed by PCA may contain badly classified elements, for this we will classify the countries by a hierarchical ascending classification. (CAH) [10]. The risk of lack of information or imperfect information is always present because of the insufficient percentage of inertia explained by the first factorial axis. Indeed, the proximity between two points on the factor map does not necessarily reflect that they have a similar behavior with respect to all the variables. To overcome this problem, a classification was made using the hierarchical Ascending classification method.

3.1 Hierarchical Ascending Classification

The first factorial plane (F1, F2, F3, F4) explains 81% of the information contained in the cloud. Thus, the projection on this plane can generate unreal superposition of the points. Consequently, the groups formed previously (during the application of the CPA) may contain badly classified elements. We will now apply the CAH, in order to improve the degree of homogeneity of the classes.

Using the Ward method, we can represent the classification in the form of a so-called hierarchical tree structure, also called a “dendrogram”. It is a statistical grouping process that will allow us to decide on the relative positions of the countries. .The progressive regrouping of the closest countries, using SPSS [13], by adopting the Ward method leads to the construction of the following hierarchical tree:



Source: Author

Figure 8: Grouping according to Hierarchical Tree and Classes

The application of the Hierarchical Ascending classification, allowed us to classify the different countries into five groups. The elements of each class:

Class 1 = {Morocco, Tunisia, Algeria, Iran, Chile}.
Class 2 = {Turkey, Egypt, Jordan, Syria}.
Class 3 = {Malaysia, Mexico, Peru, Brazil}.
Class 4 = {South Korea}.

3.2 Characteristics of Selected Classes

The Hierarchical Ascending classification, allows us to have the characteristics of selected classes:

Table 5
Characteristics of Selected Classes

Classe	YER	YUR	LE	NBI	NDI	NMUI	MS	NPRI	R&D	NTP	P.P	ER
1	-0,12	0,26	0,1	-0,3	-0,68	-0,36	0,51	0,52	-0,13	-0,29	0,49	-0,25
2	-0,48	0,213	0,1	-0,2	0,79	-0,47	0,33	-1,01	-0,56	-0,31	-0,22	-0,79
3	1,24	-0,91	0,1	-0,2	0,14	0,41	-0,86	-0,01	-0,32	-0,17	-0,39	1,35
4	-0,84	-1,22	1,1	3,5	0,61	-0,43	0,40	-0,91	3,28	3,42	0,20	1,80

Source: Author

The first class consists of countries, which have a low employment rate; at the scientific level the number of trades patents and the share of gross domestic product are still insufficient. These countries, knowing average health indices compared to other classes. On the other hand, poverty remains one of the major problems which oppose this class.

The second class is characterized by a weak sanitary system: the number of hospital beds by 10000 inhabitants' remains poor, so the number of murders is increasing compared to other classes. The youth unemployment rate remains high in this class, which represents obstacles for this class. Poverty is one of the major problems of the third class; however, this class has a high employment rate as well as average health indices. At the scientific level, the field of research and development is still slowly improving.

South Korea, which constitutes only one class is characterized by high health indices, so its health system remains incomparable compared to the other classes. The country has a capital interest in the field of research and development, which reflects the economic power they occupy between the major economic poles. On the other hand, the youth employment rate remains one of the obstacles in this country.

CONCLUSION

In this work, the first part focused on the theoretical framework of principal component analysis. Then, the second part aimed to classify the efficiency of social spending from different angles: first by analyzing the main component of the application (PCA), which showed that the first four factors explain more than 81% of the

information. This analysis made it possible to classify Morocco among the countries which marked the least efficiency in terms of social expenditure with: Algeria, Tunisia, etc., then by an ascending hierarchical classification (CAH) of the selected countries which was carried out on the basis of these social indicators in order to group the countries into four homogeneous classes, the comparison made showed that the classifications are almost identical to those obtained by the PCA.

REFERENCES

1. Abdi, H. and Williams, L.J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4), 433-459.
2. Artus, P. (2012). Ecole Nationale de la Statistique et de l'Administration Economique, Malakoff Cedex, France. In Artus, P., Guvenen, O. and Gagey, F. (Eds.) *International Macroeconomic Modelling for Policy Decisions*, (105-128), Martinus Nijhoff Publisher, The Netherlands.
3. Barthelemy, J.P., Brucker, F. and Osswald, C. (2004). Combinatorial optimization and hierarchical classifications. *Quarterly Journal of the Belgian, French and Italian Operations Research Societies*, 2(3), 179-219.
4. Dudley, R.M. (2014). *Uniform central limit theorems* (Cambridge Studies in Advanced Mathematics 142). Cambridge University Press.
5. Genesereth, M.R. and Ketchpel, S.P. *"Software Agents"*. Online available at http://staff.um.edu.mt/mmon1/lectures/csa3210/GenKet.pdf
6. Landau, S. (2004). *A handbook of statistical analyses using SPSS*. CRC.
7. Rosenblatt, M. (1956). A central limit theorem and a strong mixing condition. *Proceedings of the National Academy of Sciences of the United States of America*, 42(1), 43-47.
8. Saporta, G. and Keita, N.N. (2009). Principal component analysis: application to statistical process control. *Gérard Govaert. Data Analysis*, ISTE, pp. 1-23, 1-84821-098-1. ff10.1002/9780470611777.ch1ff. ffhal-01125713f
9. Wold, S., Esbensen, K. and Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1-3), 37-52.
10. *World Bank Open Data* are available on the World Bank website from: https: //www://donnees.banquemondiale.org
11. Zhang, B., Fu, M. and Yan, H. (2001). A nonlinear neural network model of mixture of local principal component analysis: application to handwritten digits recognition. *Pattern Recognition*, 34(2), 203-214.

APPENDIX

The R code of the Principal Component Analysis application:

```
library("FactoMineR"), library("factoextra"), library("factoextra") “libraries needed PCA.

data<read.csv("C:/Users /Desktop/PCA.csv",sep=";") “function to import of data”.

PCA (data,quali.sup=1:1,ind=1:1, graph = FALSE) “ function to calculate PCA”.

APCA<- PCA(data,quali.sup=1:1,ind=1:1, graph = FALSE) “PCA calculation on active individuals /variables”

eig.val <- get_eigenvalue(APCA) “ eigenvalues / Variances”.

fviz_eig(APCA, addlabels = TRUE, ylim = c(0, 50)) “The eigenvalue graph”.

var <- get_pca_var(APCA) “This function returns a list of elements containing the resultats”

head(var$coord) “ Coordinates for the variables”

head(var$cos2) “ Cos2 for the variables”

head(var$contrib) “ Contributions of the variables”

fviz_pca_var(APCA, col.var = "black") “function to view the variables”

fviz_pca_var (APCA, col.var = "contrib",gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),repel = TRUE) “Graphical representation in the factorial plan to the contribution)”.
```